

IJCSIS Vol. 16 No. 4, April 2018
ISSN 1947-5500

International Journal of Computer Science & Information Security

© IJCSIS PUBLICATION 2018
Pennsylvania, USA

Indexed and technically co-sponsored by :



Indexing Service

IJCSIS has been indexed by several world class databases, for more information, please access the following links:

Global Impact Factor

<http://globalimpactfactor.com/>

Google Scholar

<http://scholar.google.com/>

CrossRef

<http://www.crossref.org/>

Microsoft Academic Search

<http://academic.research.microsoft.com/>

IndexCopernicus

<http://journals.indexcopernicus.com/>

IET Inspec

<http://www.theiet.org/resources/inspec/>

EBSCO

<http://www.ebscohost.com/>

JournalSeek

<http://journalseek.net>

Ulrich

<http://ulrichsweb.serialssolutions.com/>

WordCat

<http://www.worldcat.org>

Academic Journals Database

<http://www.journaldatabase.org/>

Stanford University Libraries

<http://searchworks.stanford.edu/>

Harvard Library

<http://discovery.lib.harvard.edu/?itemid=|library/m/aleph|012618581>

UniSA Library

<http://www.library.unisa.edu.au/>

ProQuest

<http://www.proquest.co.uk>

Zeitschriftendatenbank (ZDB)
<http://dispatch.opac.d-nb.de/>

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS

International Journal of Computer Science and Information Security (IJCSIS) January-December 2018 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.

See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site

Notification: see web site

Revision: see web site

Publication: see web site

Context-aware systems

Networking technologies

Security in network, systems, and applications

Evolutionary computation

Industrial systems

Evolutionary computation

Autonomic and autonomous systems

Bio-technologies

Knowledge data systems

Mobile and distance education

Intelligent techniques, logics and systems

Knowledge processing

Information technologies

Internet and web technologies, IoT

Digital information processing

Cognitive science and knowledge

Agent-based systems

Mobility and multimedia systems

Systems performance

Networking and telecommunications

Software development and deployment

Knowledge virtualization

Systems and networks on the chip

Knowledge for global defense

Information Systems [IS]

IPv6 Today - Technology and deployment

Modeling

Software Engineering

Optimization

Complexity

Natural Language Processing

Speech Synthesis

Data Mining

For more topics, please see web site <https://sites.google.com/site/ijcsis/>

 arXiv.org  Google scholar

 SCIRUS
search engine for science

 ScientificCommons

 Scribd

 .docstoc
find and share professional documents

 BASE
Bielefeld Academic Search Engine

 CiteSeerX beta

 Computer Science
Bibliography

 DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS

 EBSCO
HOST

 ProQuest

For more information, please visit the journal website (<https://sites.google.com/site/ijcsis/>)

Editorial Message from Editorial Board

It is our immense pleasure to present the April 2018 issue (Volume 16 Number 4) of the International Journal of Computer Science and Information Security (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and digital technologies. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 10569 times and this journal is experiencing steady and healthy growth. Google statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, LinkedIn, Academia.edu and EBSCO among others.

A great journal cannot be made great without a dedicated editorial team of editors and reviewers. On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to meticulously review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status, making sure we deliver high-quality content to our readers in a timely fashion.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." We would like to thank you, the authors and readers, the content providers and consumers, who have made this journal the best possible.

For further questions or other suggestions please do not hesitate to contact us at ijcsiseditor@gmail.com.

A complete list of journals can be found at:

<http://sites.google.com/site/ijcsis/>

IJCSIS Vol. 16, No. 4, April 2018 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):



Open Access This Journal is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source.



Bibliographic Information

ISSN: 1947-5500

Monthly publication (Regular Special Issues)

Commenced Publication since May 2009

Editorial / Paper Submissions:

IJCSIS Managing Editor

(ijcsiseditor@gmail.com)

Pennsylvania, USA

Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	Dr Riktesh Srivastava [Profile] Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	Dr. Jianguo Ding [Profile] Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Engineering, Shahid Beheshti University, Tehran, Iran	Dr. Naseer Alquraishi [Profile] University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	Dr. Kai Cong [Profile] Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	Dr. Omar A. Alzubi [Profile] Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	Dr. Jorge A. Ruiz-Vanoye [Profile] Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	Prof. Ning Xu, Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	Dr . Bilal Alatas [Profile] Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	Dr. Ioannis V. Koskosas, University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	Dr Venu Kuthadi [Profile] University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	Dr. Zhihan Lv [Profile] Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	Prof. Ghulam Qasim [Profile] University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	Prof. Dr. Maqbool Uddin Shaikh [Profile] Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	Dr. Musa Peker [Profile] Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	Dr. Wencan Luo [Profile] University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	Dr. Ijaz Ali Shoukat [Profile] King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	Dr. Yilun Shang [Profile] Tongji University, Shanghai, China
Dr. Sedat Akylek [Profile] Ondokuz Mayis University, Turkey	Dr. Sachin Kumar [Profile] Indian Institute of Technology (IIT) Roorkee

Dr Basit Shahzad [Profile] King Saud University, Riyadh - Saudi Arabia	Dr. Mohd. Muntjir [Profile] Taif University Kingdom of Saudi Arabia
Dr. Sherzod Turaev [Profile] International Islamic University Malaysia	Dr. Bohui Wang [Profile] School of Aerospace Science and Technology, Xidian University, P. R. China
Dr. Kelvin LO M. F. [Profile] The Hong Kong Polytechnic University, Hong Kong	Dr. Man Fung LO [Profile] The Hong Kong Polytechnic University

ISSN 1947 5500 Copyright © IJCSIS, USA.

TABLE OF CONTENTS

1. PaperID 31031819: A Neighbourhood-Based Trust Protocol for Secure Collaborative Routing in Wireless Mobile D2D HetNets (pp. 1-9)

(1) *Aminu Bello Usman, (2) Jairo Gutierrez & (3) Abdullahi Baffa Bichi*

(1,2) *School of Engineering, Computer and Mathematical sciences, Auckland University of Technology, New Zealand*

(3) *Department of Computer Science, Bayero University, Kano Nigeria*

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

2. PaperID 31031822: Design and Analysis of Low Power Double Tail Comparator for 2-bit Fast ADC (pp. 10-14)

A. Anitha, M. Balaji, Ravishankar Kandasamy & S. K. Ragul Vijayan

Assistant Professor, ECE, Sri Shannmuga College of Engineering and Technology, Erode, Tamil Nadu

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

3. PaperID 31031824: Detecting Windows Operating System's Ransomware based on Statistical Analysis of Application Programming Interfaces (API) Functions Calls (pp. 15-22)

Abdulgader Almutairi, Assistant Professor, College of Sciences and Arts at Qassim University – Kingdom of Saudi Arabia (KSA)

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

4. PaperID 31031825: A Review on Congestion Control Approaches for Real-Time Streaming Application in the Internet (pp. 23-28)

Abhay Kumar, Department of CSE, J B Institute of Engineering and Technology, Hyderabad, India

P. V. S. Srinivas, Sreenidhi Institute of Technology and Science, Hyderabad, India.

Dr. A Govardhan, Department of CSE, Jawaharlal Nehru Technological University Hyderabad, India

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

5. PaperID 31031827: Student Alcohol Consumption Prediction: Data Mining Approach (pp. 29-42)

(1) *Hind Almayyan, (2) Waheeda Almayyan*

(1) *Computer Department, Institute of Secretarial Studies, PAAET, Kuwait*

(2) *Computer Information Department, Collage of Business Studies, PAAET, Kuwait*

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

6. PaperID 31031828: Big Data Challenges faced by Organizations (pp. 43-54)

Noureen Kausar, Rabia Saleem; Department of Information Technology, GC University, Faisalabad, Pakistan.

Sidra Amin, Department of Information Technology, GC University, Faisalabad, Pakistan.

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

7. PaperID 31031834: Enhancement of Degraded Document Images using Retinex and Morphological Operations (pp. 55-60)

*Chandrakala H. T., Research Scholar, Dept. of CSE, VTU Regional Research Center, Bengaluru, India
Thippeswamy G., Professor and Head, Dept. of CSE, BMS Institute of Technology, Bengaluru, India
Sahana D. Gowda, Professor and Head, Dept. of CSE, BMS Institute of Technology, Bengaluru, India*

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

8. PaperID 31031804: Comparative Analysis of K-Means Data Mining and Outlier Detection Approach for Network-Based Intrusion Detection (pp. 61-76)

*Joseph Panford & Lazarus Kwao
Dept. of Computer Science, KNUST, Kumasi, Ghana*

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

9. PaperID 31031807: Refactoring to Microservice Architecture (pp. 77-80)

Dr. Latha Sadanandam, Software Architect, Danske IT and Support Services India Pvt Ltd

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

10. PaperID 31031817: Breast Cancer Stage Classification on Digital Mammogram Images (pp. 81-88)

*Dr. G. Rasitha Banu (1), Fathima N Sakeena (2), Mrs. Mumtaj (3), Mr. Agha Sheraz Hanif (4)
(1) Assistant Professor, Faculty of PHTM, Dept. of HI, Jazan University, KSA
(2) Lecturer, Faculty of CS& IS, Jazan University, KSA
(3) Assistant Professor, Dept. of bioinformatics, Md. Sathak College, India
(4) Lecturer, Faculty of PHTM, Department of HI, Jazan University, KSA*

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

11. PaperID 31031818: Optimizing Building Plan for a (9m X 12m) House Using Learning Systems (pp. 89-95)

*Dr. Khalid Nazim S. A. (1), Dr. Harsha S. (2), Abhilash Kashyap B. (3), Dr. Fayez Al Fayez (4)
(1) Assistant Professor, Department of CSI, College of Science, Majmaah University, Majmaah 11952, Saudi Arabia
(2) Associate Professor, Department of ISE, JIT, Bengaluru
(3) 6th CSE, JIT, Bengaluru
(4) Assistant Professor, Department of CSI, College of Science, Majmaah University, Majmaah 11952, Saudi Arabia*

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

12. PaperID 31031826: A Survey on Rule-Based Systems and the significance of Fault Tolerance for High-Performance Computing (pp. 96-102)

*G. Sreenivasulu, Department of CSE, J B Institute of Engineering and Technology, Hyderabad, India
P. V. S. Srinivas, Sreenidhi Institute of Technology and Science, Hyderabad, India
Dr. A Govardhan, Department of CSE, Jawaharlal Nehru Technological University Hyderabad, India*

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

13. PaperID 31031836: Risk Assessment: Approach to enhance Network Security (pp. 103-110)

*(1) Nwagu, Chikezie Kenneth; (2) Omankwu, Obinnaya Chinecherem; (3) Okonkwo, Obikwelu Raphael
(1) Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
(2) Computer Science Department, Michael Okpara University of Agriculture Umudike Umuahia, Abia State, Nigeria.
(3) Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.*

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

14. PaperID 31031846: Sag-Tension Analysis of AAAC Overhead Transmission Lines for Hilly Areas (pp. 111-114)

*Muhammad Zulqarnain Abbasi, M. Aamir Aman, Hamza Umar Afridi, Akhtar Khan
IQRA National University, Pakistan*

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

15. PaperID 31031837: A Review of Intelligent Agent Systems in Animal Health Care (pp. 115-116)

*(1) Omankwu, Obinnaya Chinecherem; (2) Nwagu, Chikezie Kenneth; (3) Inyiama, Hycient
(1) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(2) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria*

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

16. PaperID 31031852: Simulated Annealing Algorithm for VLSI Floorplanning for Soft Blocks (pp. 117-125)

*Rajendra Bahadur Singh, Dept. of Electronics & Comm., School of ICT Gautam Buddha University, Greater Noida, INDIA
Anurag Singh Baghel, Dept. of Computer Science, School of ICT, Gautam Buddha University, Greater Noida, INDIA*

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

17. PaperID 31031838: A Review of Expert Systems in Agriculture (pp. 126-129)

*(1) Nwagu, Chikezie Kenneth; (2) Omankwu, Obinnaya Chinecherem; (3) Inyiama, Hycient
(1) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria
(2) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria*

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

18. PaperID 31031853: Keywords- Based on Arabic Information Retrieval Using Light Stemmer (pp. 130-134)

*Mohammad Khaled A. Al-Maghasbeh & Mohd Pouzi Bin Hamzah,
School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia*

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

19. PaperID 31031854: Kinect Sensor based Indian Sign Language Detection with Voice Extraction (pp. 135-141)

Shubham Juneja (1), Chhaya Chandra (2), P. D. Mahapatra (3), Siddhi Sathe (4), Nilesh B. Bahadure (5) & Sankalp Verma (6)

(1) Material Science Program, M.Tech first year student, Indian Institute of Technology, Kanpur, UttarPradesh, India

(2) B.E, Electrical and Electronics Engineering, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

(3) B.E, Electrical and Electronics Engineering, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

(4) Department of Electrical & Electronics Engineering, B.E. Final Year Student, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

(5) Assoc. Professor at Department of Electronics & Telecommunication Engineering, MIT College of Railway Engineering & Research, Solapur, Maharashtra, India

(6) Assoc. Professor at Department of Electrical & Electronics Engineering, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

20. PaperID 31031839: Computer Science Research Methodologies (pp. 142-144)

(1) Omankwu, Obinnaya Chinecherem; (2) Nwagu, Chikezie Kenneth; (3) Inyiamma, Hycient

(1) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria

(2) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,

(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

21. PaperID 31031859: Recovery of RGB Image from Its Halftoned Version based on DWT (pp. 145-150)

Tasnim Ahmed, Department of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh

Md. Imdadul Islam, Department of Computer Science & Engineering, Jahangirnagar University, Dhaka, Bangladesh

Md. Habibur Rahman, Department of Computer Science & Engineering, Jahangirnagar University, Dhaka, Bangladesh

[Full Text: PDF](#) [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

22. PaperID 31031862: Data Redundancy on Diskless Client using Linux Platform (pp. 151-154)

B. S. Sonawane, Research Fellow, Dept. of CSIT, Dr. B. A. M. University, Aurangabad

R. R. Deshmukh, Professor, Dept. of CSIT, Dr. B. A. M. University, Aurangabad

S. D. Waghmare, Research Fellow, Dept. of CSIT, Dr. B. A. M. University, Aurangabad

Pushpendra Chavan, Principal, Tech Support Engineer, Red Hat India Pvt. Ltd, Pune,

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

23. PaperID 31031840: Enhanced Feature Analysis Framework for Comparative Analysis & Evaluation of Agent Oriented Methodologies (pp. 155-159)

(1) Omankwu, Obinnaya Chinecherem; (2) Nwagu, Chikezie Kenneth; (3) Inyama, Hycient

(1) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria

(2) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,

(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

24. PaperID 31031863: Suitability of Addition-Composition Fully Homomorphic Encryption Scheme for Securing Data in Cloud Computing (pp. 160-173)

Richard Omollo & George Raburu

Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, P.O. Box 210-40601, Bondo-Kenya

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

25. PaperID 31031871: Conversion Prediction for Advertisement Recommendation using Expectation Maximization (pp. 174-183)

Sejal D., Department of Computer Science and Engineering, B N M Institute of Technology, Bangalore

Shradha G. & Venugopal K. R., Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore-560001

S. S. Iyengar, Florida International University, USA

L. M. Patnaik, INSA Senior Scientist, National Institute of Advanced Studies, IISc Campus, Bangalore

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

26. PaperID 31031872: An Android Application Studhelper for Engineering Students (pp. 184-187)

Ishani Mukherjee (1), Aman Bansal (2), Mokshada Patra (3,) Rahul Pal (4), Md. Khaja Mohiddin (5)

UG Scholars (1,2,3,4), Senior Assistant Professor (5)

Department of Electronics & Telecommunication Engineering, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

27. PaperID 31031874: Live Forensics Analysis Method for Random Access Memory on Laptop Devices (pp. 188-192)

(1) Danang Sri Yudhistira, (2) Imam Riadi, (3) Yudi Prayudi

- (1) Department of Informatics, Universitas Islam Indonesia
(2) Department of Information System, Universitas Ahmad Dahlan
(3) Department of Informatics, Universitas Islam Indonesia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

28. PaperID 31031882: Evaluation of Snort using Rules for DARPA 1999 Dataset (pp. 193-199)

Ayushi Chahal & Dr. Ritu Nagpal
Department of Computer Science and Engineering, Guru Jambheshwar University of Science & Technology, Hisar, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

29. PaperID 31031884: A Low Cost ECG Monitoring System with ECG Data Filtering (pp. 200-204)

Md. Rakib Hasan, Dept. of CSE, Jahangirnagar University, Dhaka, Bangladesh
Rabiul Alam Sarkar, Dept. of CSE, Jahangirnagar University, Dhaka, Bangladesh
Md. Firoz-Ul-Amin, Dept. of CSE, Jahangirnagar University, Dhaka, Bangladesh
Mohammad Zahidur Rahman, Dept. of CSE, Jahangirnagar University, Dhaka, Bangladesh

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

30. PaperID 31031885: The Convenience Activity on TQM of Advanced Technology on User Expectation of Online Banking Systems in India (pp. 205-213)

Mohd Faisal Khan & Dr. Debaprayag Chaudhuri
Department of Information Technology (IT), AMET University Chennai, Tamil Nadu (India),

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

31. PaperID 31031886: Feasibility Analysis of Directional-Location Aided Routing Protocol for Vehicular Ad-hoc Networks (pp. 214-225)

Kamlesh Kumar Rana, Computer Science & Engineering, IIT (ISM) Dhanbad, Jharkhand, India
Sachin Tripathi, Computer Science & Engineering, IIT (ISM) Dhanbad, Jharkhand, India
Ram Shringar Raw, Computer Science & Engineering, IGNTU Amarkantak, Madhya Pradesh, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

32. PaperID 31031897: San Bernardino Symphony Orchestra and Exploring the Use of Mobile Applications by Symphony Orchestras (pp. 226-250)

Abdullah Almusallam, Evelia Avila, Prabhjeet Grewal, Abdulmajid Alnoamani

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

33. PaperID 310318101: Improved Text Mining for Bulk Data Using Deep Learning Approach (pp. 251-254)

Indumathi A., PG Scholar, Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore.

Perumal P., Professor, Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

34. PaperID 310318103: A Ranking Model for Software Requirements Prioritization during Requirements Engineering; A Case Study (pp. 255-268)

Ishaya P. Gambo, Department of Computer Science and Engineering, Faculty of Technology Obafemi Awolowo University, Ile-Ife, Nigeria

Rhoda N. Ikono, Department of Computer Science and Engineering, Faculty of Technology Obafemi Awolowo University, Ile-Ife, Nigeria

Philip O. Achimugu, Department of Computer Science, Lead City University, Ibadan, Nigeria

Olaronke G. Iroju, Department of Computer Science, Adeyemi College of Education, Ondo, Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

35. PaperID 310318105: Segmentation of Diffusion Tensor Brain Tumor Images using Fuzzy C-Means Clustering (pp. 269-272)

Ceena Mathews, Department of Computer Science, Prajyoti Niketan College, Pudukad, Thrissur, Kerala, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

36. PaperID 310318106: Partial Discharges using Variable Frequency PRPDA Technique (pp. 273-278)

M. Zubair Bhayo (1), M.Ali (2), Kalsoom Bhagat (3), Abdul Hameed (4)

(1,2,3) Department of Electrical Engineering, Mehran UET SZAB Campus Khairpur Mir's,

(4) School of Automation, Northwestern Polytechnical University Xian China

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

37. PaperID 28021829: Vehicle Power Line Channel Modelling under CST Microwave Studio (pp. 279-282)

Mohammed Fattah, Transmission and Information Processing Team, Moulay Ismail University, Meknes, Morocco

S. Mazer, M. El Bekkali, R. Ouremchi, M. El Ghazi

Transmission and Information Processing Laboratory, Sidi Mohamed Ben Abdellah University, Fez, Morocco

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

38. PaperID 28021849: Big Data Clustering Model based on Fuzzy Gaussian (pp. 283-288)

Amira M. El-Mandouh, Beni-Suef University

Laila A. Abd-Elmegid, Helwan University

Hamdi A. Mahmoud, Beni-Suef University

Mohamed H. Haggag, Helwan University

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

39. PaperID 28021852: Cross Layer Based Hybrid Fuzzy Ad-Hoc Rate Based Congestion Control (CLHCC) Approach for VoMAN to Improve Quality of VoIP Flows (pp. 289-297)

*V. Savithri, Assistant Professor, Part-Time Ph.D. Scholar, Bharathiar University, Coimbatore, India
Dr. A. Marimuthu, Associate Professor & Head, PG and Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

40. PaperID 31121709: Caesar Cipher Method Design and Implementation Based on Java, C++, and Python Languages (pp. 298-307)

Ismail M. Keshta, Department of Computer and Information Technology, Dammam Community College (DCC), King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

41. PaperID 310318109: Chronic Kidney Disease Prediction Using Machine Learning (pp. 308-311)

Sathiya Priya S., PG Scholar, Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore

Suresh Kumar M., Professor, Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

42. PaperID 31031849: Learning Engagement Based on Cloud Classroom (pp. 312-319)

*Nasrin Akter, Shah Akbar Ahmad
Computer Science and Engineering, Daffodil International University*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

43. PaperID 31031857: Self Organizing Migration Algorithm with Curvelet Based Non Local Means Method for the Removal of Different Types of Noise (pp. 320-330)

*Sanjeev K. Sharma, Associate Professor, Department of E&I, SATI, Vidisha (M.P.)
Dr. Yogendra Kumar Jain, Professor and I/C HOD, Department of CSE, SATI, Vidisha (M.P.)*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

44. PaperID 31101607: Discovery of Jumping Emerging Patterns Using Genetic Algorithm (pp. 331-336)

*Sumera Qurat ul Ain, Saif ur Rehman
UIIT, Arid Agriculture University, Rawalpindi, Pakistan*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

A Neighbourhood-Based Trust Protocol for Secure Collaborative Routing in Wireless Mobile D2D HetNets

Aminu Bello Usman ^{#1}, Jairo Gutierrez ^{*2} Abdullahi Baffa Bichi ^{#3}

^{#1,*2} School of Engineering, Computer and Mathematical sciences
Auckland University of Technology, New Zealand

¹ ausman@aut.ac.nz

² jairo.gutierrez@aut.ac.nz

^{*3} Department of Computer Science
Bayero University, Kano Nigeria

³ ab'affa.cs@buk.edu.ng

Abstract—Heterogeneous Device-to-Device mobile networks are characterised by frequent network disruption and unreliability of peers delivering messages to destinations. Trust-based protocols has been widely used to mitigate the security and performance problems in D2D networks. Despite several efforts made by previous researchers in the design of trust-based routing for efficient collaborative networks, there are fewer related studies that focus on the peers' neighbourhood as a routing metrics' element for a secure and efficient trust-based protocol. In this paper, we propose and validate a trust-based protocol that takes into account the similarity of peers' neighbourhood coefficients to improve routing performance in mobile HetNets environments. The results of this study demonstrate that peers' neighbourhood connectivity in the network is a characteristic that can influence peers' routing performance. Furthermore, our analysis shows that our proposed protocol only forwards the message to the companions with a higher probability of delivering the packets, thus improving the delivery ratio and minimising latency and mitigating the problem of malicious peers (using packet dropping strategy).

Heterogeneous Networks, Wireless D2D Communications, Trust-Based Protocol, Secure Collabaorative Routing.

I. INTRODUCTION

The Heterogeneous (HetNets) Device-to-device (D2D) networks that enable direct communication between nearby mobile devices is an exciting innovation that facilitates interoperability between critical public safety networks and increases the amount of traffic, quality requirements, and enables new mobile cloud computing demands. Among others, one important feature of D2D communication is the direct communication with the immediate next-hop peer, which in turn, have several advantages such as increasing network spectral efficiency and energy efficiency, and reducing transmission delay. Along with these advantages, the D2D in the heterogeneous network has various anticipated challenges for collaborative routing. For example, a device may or may not cooperate in data forwarding; a device may fail to appropriately participate in collaborative task due to its limited resources, or position in the network. Also, some routing protocols of D2D com-

munications use the assumption, that in a cooperative HetNets environments, a peer with higher trust value can serve as a good potential relay regardless of the peer's connectivity, and the number of neighbours at a time. This, however, may not be a valid supposition in practice. For example in a Delay tolerant network (DTN), a peer with a good record of data forwarding may receive a packet but can fail to forward it on time if there are no immediate neighbours around or if its neighbours are not connected with other peers that can forward the messages to the destination.

Additionally, the ability of the peers in wireless communication settings to transmit the data packet across the network is limited to the proximity between peers' communication ranges and the energy applied by peers when sending data. In such environments, two peers can communicate with each other only when they are in contact (in the same transmission range with each other). That is to say, a peer that is in a strategic location (connected) can have a higher probability of transmitting the data packets across the network.

Thus, the success of collaborative routing in HetNets depends on the extent at which the peers can fully interact with other peers and peers can make a routing decision based on trust, cooperation, and indeed the level of peers' connectivity in the network. The resultant collaborative routing task between the peers empower the peers to engage in greater tasks beyond those that can be accomplished by individual peers in the network [1], and it helps the peers in making collective routing decisions and judgement about the behaviour and actions of other peers in the network. In fact, collaborative routing between the peers improves the D2D Wireless Network efficiency [2] and enables efficient packet routing and data forwarding. At the same time it prevents jamming and minimises end-to-end delay and latency [3] as well as improving the data-centric behaviour of many WSNs applications [4]. In collaborative routing schemes, a peer may altruistically contribute their resources or serve as a good relay peer for the satisfaction of being an active

contributor or to gain recognition (increase in popularity or trust level). Also, peers can collaborate and cooperate in the processes of traffic relaying, outlier analysis, or next neighbour selection to maximize total network throughput by using all the available peers for routing and forwarding. This perception made it clear that the more the peers participate positively in the routing processes, the higher the network performance, and the higher the chances for the network to be secured regarding the denial of service attacks (Sybil attacks, blackhole attacks, etc.). However, the collaborative routing mechanism along with its advantages brings in some security issues such as information errors and losses caused by components' failure of peers in the network, external interference, wireless transmission errors and excessive packet drops [5] which can adversely affect the delivery performance of data communication in the network. All these challenges might be related due to the peers' inability to identify an excellent relaying peer while making a routing decision. Therefore, the success of collaborative routing mechanisms used by wireless D2D devices largely depends on the extent at which the peers can make efficient routing decisions through identifying a connected peer that can serve as a good relaying peer in the network. For example, in many D2D networks such as MANET, WSN, VANETs etc. peers are expected to utilize their limited resources (energy) for routing functions (next peer selection, data forwarding etc.) with the probability of higher packets deliverance.

Previous studies have shown that the cooperation, and collaboration enforcement mechanism between the peers using trust and reputation, can increase the network performance and quality of service [6],[7] in the network. Currently, many D2D trusts and reputation models have been proposed in different kinds of literature. Most of the proposed trust and reputation models only consider the models' implementation based on the satisfactory and unsatisfactory behaviour of the peers in providing the valid packets [8] and the use of community-based reputations to estimate the trustworthiness of the peers (peer trust models) [9] to mention but few. However, the modelling of trust and reputation in D2D networks is a critical mission that cannot be accomplished without considering the information processing and communication across the networks and the connectivity between the peers for effectively distributed trust decision making. In this regards, is of great interest to understand how the peers' contacts and connectivity can influence trust decision processes and by extension can influence routing performance. Here, we note that with the short-range wireless transmission ability of the mobile peers, is possible in the absent of global network connectivity, that the mobility pattern of the peers can be exploited for packet transfer between the peers, even though, there might be no end-to-end connectivity.

Further, recent studies including the work of [10] shows that the network of Mobile HetNets exhibits a ubiquitous of Transient Connected Components (TCCs).That peers make

contacts and interact with other peers to form connected components which can enable peers to contact each other through multi-hop wireless connection. Further, some studies including the work of [11] have shown that the use of social metrics and Complex Network Analysis (CNA) such as peers' Centrality Estimation for computing the comparative centrality of two encountering peers and similarity of the peers' behavioural profiles based on the mobility preferences [12], peers' betweenness and community structure [13] can be exploited to provide effective solution to improve the performance of routing forwarding between mobile peers. Further, previous studies suggested that the peers' mobility can equally contribute to predicting peers' contacts, peers connectivity and peers ability in delivering the packets from the sources to the destinations [14]. We understand that the community structure is an essential properties of CNA that reveals the inherent structure of the complex network and can be used for predicting the future contacts in mobile networks such as DTNs MANETs etc [13].

One of the basic measures to describe the peers' network connectivity is the distribution of the number of links (established wireless connections with other neighbours) per network-node and the number of shared neighbours among the peers. However, to investigate peers' connectivity of mobile network settings in relation to routing between the peers, it is essential to understand the basic principles behind peers, contacts, peers' mobility pattern and a key step in establishing contact between peers among the neighbours (discovery process).

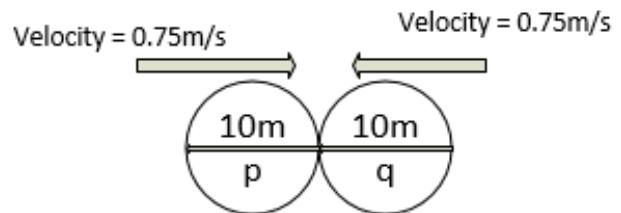


Fig. 1. Contacts Illustration

For example, consider the diagram in Figure 1, which illustrates the encounters between the two mobile peers, p and q who move in opposite directions (toward each other), with each peer having a diameter range of 10m moving at a velocity of 0.75m/s. It can be noticed that the contacts and connectivity between the peers depend on three factors. The first factor is the diameter of the peers (wireless range covered); with a diameter = 10m, it will take the peers a minimum window contact opportunity of only 26.7s to overlap each other (to go out of range of each other), and when the diameter of the devices is 15m each, and with the same velocity of 0.75m/s, the contact opportunity can be up to 40s. Here, we note that the contact time is the time it takes the peers to discover each other and establish a communication channel. Although this

is trivial, it can be noticed that the devices' wireless ranges influence the devices' contact duration and connectivity.

The second factor is the speed of the devices. With the increase of the peers' speed, the minimum contact time decreases. The third factor which is related to our interest is the frequency of contacts between the peers (this will be covered in detail in subsequent sections).

However, some of the related questions for understanding the connectivity between peers in relation to forwarding metrics include: how can the peer's contacts be appropriately captured and represented adequately in the neighbours' discovery process for trust evaluations? Can the peers' contacts, motions serve as a basis for peers' connectivity and peers' ranking in trust evaluation metrics? How can the decision trust be integrated with the peers' connectivity ranks for collaborative routing decisions? To attempt these questions, we contribute in the following ways:

(i) We propose a neighbourhood trust-based data forwarding strategy to improve the performance of wireless mobile devices in a HetNets. We achieved this through developing a similarity algorithm for quantifying the peer's similarity regarding the number of neighbours as a forwarding metric.

(ii) We propose a new trust-based protocol for evaluating peers routing behaviour.

(iii) We analyse the relationship between peers' connectivity, peers' radio ranges, and peers' speed for a trust-based routing decisions as a new way of understanding peers' attributes as elements of trust evaluation between wireless peers.

For the experimentation, we observed performance of our proposed solution using the Opportunistic Networking Environment (ONE) Simulator [15].

II. RELATED WORK

Several trust-aware models based on routing and resource attributes for peer selection using measures of trustworthiness and peers routing abilities were proposed in the literature[16]. There are few efforts from the literature that explore the influence of peers' connectivity as a trust evaluation factor in D2D collaborative routing schemes[17]. The use of social properties such as friendship [18], community structure [19], similarities regarding peers' interests [20] and location of peers [21] has recently become the focus of many collaborative routing schemes. Yet, there has been less work dedicated to clearly establish the relationship between different peers' routing elements such as peers' speed, interface range and peers' overlay connectivity in relation to peers routing reliability for trust-based routing protocols.

Further, many of the ad-hoc network trust models are naively based on a trust-your-neighbour relation. In this type of trust model, the entire trust management system (origination, managing and expiration) usually has a short lifespan, and the peers may lack a comprehensive knowledge of the overall neighbour trust level. As a result, most of the direct trust models only work in an environment where all the nodes are self-organized and mobile (e.g., military and law enforcement

applications) which limits their functionalities to some specific areas.

Recently, several attempts were made by many authors to propose a different improvement in the various aspects of direct trust and reputation algorithms. For example, the study in [22] introduced a zone-based trust management agreement scheme in wireless sensor networks. The scheme was designed to detect and revoke groups of compromised nodes within the trust formation and forwarding phase. Each node directly interacts with the neighbouring nodes for the trust report event and stores the report in a knowledge cache. The proposed protocol comprises of zone discovery, trust formation and forwarding phases. Before making a final judgement, a trustor will always compute the difference between the probability distribution function of the neighbourhood trust and the probability distribution function of the information received from its neighbours at every slot of time (say, T). The total trust factor can be determined based on the deviation between the reports of the observation using the information theoretic metric KullbackLeibler-divergence.

Also, the work in [23] proposed a novel, Connected Dominating Set (CDS)-based reputation monitoring system. Which employs a CDS-based monitoring backbone to securely aggregate the reputation of sensors without subjecting them to energy depletion or reputation pollution attacks.

In addition, apart from constraints that are application-specific, the concept of direct trust suffers from the following setbacks that may limit its application in a distributed and autonomous wireless network: a) Notion of prediction: peer p can either trust peer q or distrust peer q [24], since it has no other means of trusting peer q ; b) peer p can only compute peer q 's trust value under the condition that peer p trusts peer q ; c) Energy depletion problem; the amount of energy needed for a wireless node to accomplish trust management processes (trust aggregation and trust evaluation) with all other neighboring peers in a distributed network will be high, since the trust between peers can only be derived based on their direct contacts and the energy needed for the node to communicate with other peers is proportional to its distance with other peers in the network [25].

III. NEIGHBOURHOOD CONNECTIVITY MODEL

Connectivity and community structure recently became the central focus of behaviour-oriented and opportunistic routing paradigms and delay tolerant networks [26]. Recently, some researchers incorporated Complex Network Analysis (CNA) to formulate and predict the future contact between peers and the peers' reliability and relay selection strategy [27]. The community structure is one of the most important properties of Complex Network Analysis. In a simple terms, a network is said to have community structure if the peers of the network can be easily grouped together into (potentially overlapping) sets of peers and each peer in the network can efficiently interact with other peers either through direct contact or indirect contacts. Based on the previous findings in the literature [17], evidence suggests that to improve routing performance

between wireless peers, taking advantage of positive social characteristics such as community structure and friendship to assist packet forwarding is essential. Additionally, the concept of community structure in relation to transitivity also goes in line with the dynamic balance theory and Simmelian triangle theory which states: "The localised cohesion between transitive peers is optimal for sharing information, encouraging cooperation and minimising conflict between the actors" [28]. The tendency for a peer to belong to a certain structured community with a similar neighbourhood can represent its potential reliability to handle a particular task to improve the quality of communication in the network and serve as a relay peer in dealing with the task of packet forwarding. Motivated by the different social network and delay tolerant routing protocols that use the history of the encounter between the peers and the transitivity in estimating each peers' delivery probability. Therefore, due to the uncertainty in nodal mobility, we foresee that identifying a particular node that belongs to a community within an arbitrary mobile wireless network can provide a new angle of view in the design of trust-based routing protocols. Thus, to understand the effects of connectivity as routing attributes, we identify a neighbourhood coefficient as a measure of the degree to which peers in the network tend to cluster together.

Given a network $G = (N, L)$ with peers' sets N and the links between the peers L . Each peer in the network can be a source or destination of the traffic, and with equal transmission range $\iota(n)$. We can define the network as $G = (N, L) : N = \{p, q, \dots, r\}$ and $L \subseteq \{(p, q)\} : p, q \in N$, and $p \neq q$. Let the transmission range of peer p be $\iota(p)$ and the distance between peer p and peer q be $dis_{p,q}$. Let n_p denotes the set of peers that are neighbours of peer p and within the communication area of p . For the communication between peer p and peer q to be successful the following condition must be satisfied: (i) $dis_{p,q} \leq \iota(n)$ (receiver is within the communication range of the sender) and any peer r such that $dis_{r,q}\iota'(r)$, is not transmitting (i.e, the receiver is free of interference from any other possible sender). In other words, peer p can successfully transmit the packet to q if p is a neighbour of q and no other q 's neighbour is transmitting to peer q simultaneously.

A. Neighbourhood Coefficient Approximation

Given a network $G = (N, L)$ consisting of peers $N = \{p, q, r\}$ and the set of communication links between the peers $L \subseteq \{(p, q)\} : p, q \in N$, and $p \neq q$, the set of neighbourhood of peer p is defined as its immediately connected neighbours as follows $N_p : N_p = \{q : \{(p, q)\} \in L \wedge \{(q, p)\} \in L\}$. If $\{(p, q)\}$ is distinct from $\{(q, p)\}$, for each peer $p \in N$ there are possible number of distinct wireless interface connection $n_p(n_p - 1)$ that could exist among the peers within the neighbourhood of peer p , where n_p is the total number of peer p neighbours.

Therefore, if we denote the neighbourhood coefficient of peer p as $N_{coef(p)}$, we can compute the routing metric of peer p using the following clustering coefficient equation (1).

$$N_{coef(p)} = \frac{2\{\{(p, q) : p, q \in N_p, \{(q, p)\} \in L\}\}}{(n_p(n_p - 1))} \quad (1)$$

Figure 2 shows an example of a community neighbourhood network of nine peers with their corresponding neighbourhood coefficient in Table I using equation (1). Suppose that peer p have a data to forward to the destination in the form of "store-carry-and-forward". Based on the existing trust mechanism in the literature, peer p can forward the packets to a peer with the higher trust value among y, f and r whose have a direct contact (indicated line between the peers) with peer p . We used the clustering coefficient equation (1) to compute the peer's neighbourhood coefficient of Figure 2 to arrived at Table I. Looking at Table I, since peer r has higher forwarding metrics regarding connectivity, it may have a higher chance of routing to the destination. Thus, it might be additional routing intelligence if peer p can evaluate its subjects' neighbourhood coefficient as an additional element of trust evaluations.

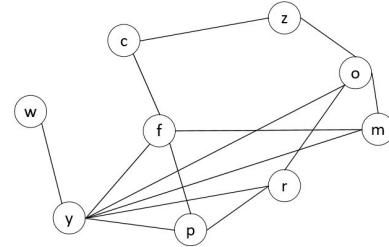


Fig. 2. Neighbourhood Illustration

TABLE I
FORWARDING METRICS TABLE

Peers	$N_{coef(p)}$	$N.Pairs = \frac{n_p(n_p - 1)}{2}$
w		0.000
z	0.000	1.000
c	0.000	1.000
o	0.333	6.000
y	0.333	15.000
m	0.667	3.000
f	0.333	6.000
r	0.667	3.000
p	0.667	3.000

B. Neighbourhood Similarity Modelling

The similarity in terms of frequent contacts, visited locations or interests are often seen as major factors for connectivity in DTN and social networks. For instance, people tend to connect with those sharing similar tastes, social background, interests and beliefs, and also similar popularity. This is often expressed as "love of the same" or "Birds of a feather flock together" [29]; that is the tendency of individuals to associate and bond with similar others which can be treated synonymously with similarity in the context of connectivity. On the same vein, peers' similarity as a network formation model can also reproduce the commonly perceived power-law or scale

free distribution of sparsely connected networks. It should be clear that in a traditional random network the degrees of all peers are distributed around the average [30], therefore using similarity of peers degree can equally be applied in collective or collaborative routing for behaviour analysis and outlier analysis for networks anomaly identification. Subsequently, a similarity model can produce the characteristics of different densities in real networks, thus, it can be used as a model for describing the topological transition between the peers in the network [31]. Therefore, the tendency of a peer to belong to a certain structured community with a similar neighbourhood can represent its potential routing ability for data forwarding within the community thus, will improve the quality of communication in the network [32]. Additionally, motivated by different social networks and delay tolerant routing protocols [33] that use the history of encounters and similarity between the peers in terms of frequent visited locations and mobility patterns for predicting peers' delivery probability, we proposed a connectivity similarity model. In this regard, we postulate that the relative comparisons of the proliferation of peers' transitivity coefficients may give a meaningful basis for understanding the peers' connectivity for collaborative routing handling [34]. We envision that such comprehensive approach has two advantages:

1. It stimulates behaviour-aware message routing protocol thereby each peer can understand the changes of its potential routing partners' connectivity; thus determining whether a peer will be a good relaying peer or otherwise.

2. It also speeds up the discovery of the peer with similar behaviour and mobility pattern for collaborative routing decisions.

Thus each time peer p wants to participate in a routing process, it will advertise its neighbourhood coefficient. This can be achieved through simple scanning of its neighbours.

$$d(N_{coef(p)}, N_{coef(q)}) = |N_{coef(p)} - N_{coef(q)}| \quad (2)$$

We can, therefore, normalise the difference between the two possible attributes values with the maximum possible attributes level as follows.

$$d(N_{coef(p)}, N_{coef(q)}) = \sum_{i=1}^n \frac{|N_{coef(pi)} - N_{coef(qi)}|}{\text{Max}(N_{coef(pi)}, N_{coef(qi)}),} \quad (3)$$

Therefore, the similarity between peer p and q 's neighbourhood coefficients can be evaluated as: $S_{p,q} = 1 - d(N_{coef(p)}, N_{coef(q)})$ which can be represented as follows:

$$S_{p,q} = 1 - \sum_{i=1}^n \frac{|N_{coef(pi)} - N_{coef(qi)}|}{\text{Max}(N_{coef(pi)}, N_{coef(qi)})} \quad (4)$$

From equation (1), the $\text{Max}(N_{coef(pi)}, N_{coef(qi)}) = 1$, therefore, the similarity between peer p and peer q is simply $1 - \sum_{i=1}^n |N_{coef(pi)} - N_{coef(qi)}|$. We can simplify the similarity of peers' neighborhood coefficient using the following equation (5).

$$S_{p,q} = 1 - |N_{coef(p)} - N_{coef(q)}| \quad (5)$$

IV. TRUST MODEL

In this section, we describe the trust evaluations between mobile peers. Upon an encounter between two peers (p, q) , peer p can update its direct trust on peer q based on the update of the total. For example, let $t_{p,q}$ be the trust value that peer p places in peer q based on its a priori experience with peer q , where $t_{p,q} \in \langle 0, 1 \rangle : p \neq q$. Looking at the distributes *EigenTrust* algorithm [8], each time peer p encounters peer q , peer p can assess the trust level of peer q based on their encounter delivery vectors exchanges. If the encounters history is not satisfactory it will be considered as a negative experience, therefore the local trust value ($t_{p,q}$) between p and q will decrease; while if the encounter history between the peers is satisfactory, then it will be considered a positive experience and the ($t_{p,q}$) will increase. If the peers' transaction is undecided, it will have no effect in the peers' trust evaluation. Therefore, $sat(p, q)$ represents the number of satisfactory encounters between peer p and peer q while $unsat(p, q)$ represents the total number of unsatisfactory encounters between peer p and peer q [35]. Evidence of trustfulness is manifested by the encounters history exchanges between the peers. Thus, the resultant local trust value between the peers can be computed as $C_{p,q} = sat(p, q) - unsat(p, q)$. The normalised reputation can be computed as:

$$t_{p,q} = \frac{\max(C_{p,q}, 0)}{\sum_q \max(C_{p,q}, 0)}, \|\vec{t_p}\| : \sum_{q=1}^N t_{p,q} = 1 \quad (6)$$

The global trust equation peer p can estimate about peer r based on the feed back of peer q about the behaviour of peer r can be presented in the following equation (7).

$$T_{p,q} = \sum_q t_{p,q} t_{q,r} \quad (7)$$

Therefore, each peer will maintain the local trust observation vectors of its subjects' trust values as follows:

$$\vec{t_p} = (t_{p,q}, \dots, t_{p,N})^T, 0 \leq t_{p,q} \leq 1 \quad (8)$$

Note: the local trust value ($t_{p,q}$) in equation (8) represents the normalised local trust value peer p have about q and other peers in the network; $T_{p,r}$ in equation (7) is global (transitive trust) of q computed by p based on trust that p has about q . Therefore, every peer can use his global observation vector's elements ($\vec{t_p}$) to compute the global trust value $T_{p,q}$. To secure the implementation of our protocol, we estimate the trust value of the peers to two basic principles; (1) the trust value of a peer is computed in a distributed fashion. Thus a peer does not have access to its trust information where it can be subject to alterations and (2) the trust value of a peer is computed by more than one peer so that malicious peers cannot succeed in white washing attacks.

The above presented algorithm is used to determine the trust worthiness of a peer in delivering received messages.

For instance, peer p can assess peer q 's unhealthiness based on evidence manifested due to malicious attacks detected which including packet dropping, self-promoting, bad-mouthing and ballot stuffing attacks through the encounter history exchanged from peer q . In the event where the encounter history is satisfied (e.g., using encounter tickets as in [36]), this is considered as a positive experience which can cause an increase in the trust level of q in the eyes' of p , otherwise it is considered as a negative experience which can lead to the decrease in the trust level of q .

Based on peer p 's experience about q , peer p can store the trust value of peer q and the neighbourhood coefficient similarity value with peer q : $(T_{p,q}, S_{p,q})$ after every contact between the two peers. If peer p has not stored $(T_{p,q}, S_{p,q})$, the trust value between the two peers is assumed to be zero therefore, the trust value can be recalculated at each opportunistic encounter according to the following rules:

- 1) All peers enter the mode where they can search for their neighbours using their shortest range receivers.
- 2) On finding one or more peers within the transmission range a peer can search its contact list to find the trust value of a peer and compute the corresponding neighbourhood similarity with the encountered peer before data transfer.
- 3) Every peer keeps its neighbourhood list and their corresponding trust values and adds itself as a member.
- 4) Peers keeps on exchanging their neighbourhood list.

Therefore, we define a specific function $TS(T_{p,q}, S_{p,q})$ as the resultant trust value between the peers as follows in equation (9).

$$TS = T_{p,q} * S_{p,q} \quad (9)$$

Our proposed connectivity trust model enables a peer to route a data packet to the corresponding neighbour with the higher probability of delivering the data packet to the destination. As mentioned earlier, the neighbours (connected peers) of the forwarding peer are those that are in the same transmission range with the forwarding peer with each node having a unique identifier. Once a peer is in the position to forward the data packet, it will look into its routing list for the computation of the trust value of its neighbours. The inputs to the routing decision depend on both the trustworthiness of a peer and the similarity of the peers' connectivity. This is to enable us to explore the relative effect of the peers connectivity in terms of routing handling. Once a peer aggregates all the trust values it will then filter peers' trust values and rank them before forwarding to the routing engine for a routing decision. A peer will select an optimal next-hop node from its neighbours using the resultant computed trust values.

Based on the presented trust model, one can observe that in an ideal collaborative wireless mobile environment, all peers can choose a next-peer for routing based on their similarity in terms of connectivity and the trustworthiness of a peer in terms of reliability for routing handling; in that way a routing path can be optimised based on the peers' trustworthiness and

peers' connectivity; thus a simple connectivity trust-based protocol is achieved.

V. PROTOCOL IMPLEMENTATION

We understand that DTNs, posses most of the characteristics of wireless D2D HetNets. Therefore, throughout this paper, we consider the characteristics of DTN networks for presenting our concept. However, our proposed model can be applied to different scenarios and related applications.

To avoid routing loops, we consider a three-hop counts routing mechanism. i.e, the maximum number of hop-counts a packet can visit is limited to only 3 hops between the source and the destination. We achieve this through configuring the TTL (Time-to-Live) value of the packets so that as the packet move between hops, the packet's TTL field is decrease by one. In the event where the TTL value reaches 0, the packet is dropped by the relaying peer that decrease the value from 1 to 0.

We conducted simulations using Opportunistic Network Environment Simulator [15] which is a DTN simulator popularly known for modelling the behaviour of store-carry-forward networks [37], [38]. We assume peers' discovery takes only 40 seconds and the packet transfer depends on the resources' availability of buffer, energy, bandwidth, and Time To Leave (TTL) etc. A peer is choose as a message carrier if its trust value is higher in comparison with other peers and its connectivity with other peers is similar to that of the sending peer. A peer must also posses the trust threshold i.e., a minimum trust level required for a peer to participate in collaborative routing. At the initial stage, we implemented our trust model with the pre-trusted peers percentage, pre-trusted peers weight and zero trust node selection probabilities as presented in Table II.

A. Protocol Evaluation

We first seek to understand how the peers connectivity influences the peers' trust evaluation in mobile wireless environments. We thus simulate the peers' mobility pattern at varying speeds of $0.5m/s$ and $0.75m/s$ as shown in Figures 3, 4, Figure 5. In our simulation, as the peers move around, they keep having contacts with other peers and interact (establish contacts, exchange messages, etc). The level of peers interaction determine its connectivity which by extension influences the performance of the routing protocol[39]. Subsequently, we model the average percentage of distinct neighbours encountered both directly and through indirect contacts. For the indirect connectivity case, we mean the peers that are reachable via multi hop relay through neighbours' neighbours as in [40].

The graph of Figure 3 shows that the average distinct number of peers connected. The plot shows as the radio ranges of peers (interface range) and movements speed increases, the peers' direct and indirect connectivity of peers increases as well. The results revealed that the faster-moving peer have

TABLE II
IMPLEMENTATION PARAMETERS

Number of host	50
Number of interface per peer	2
Movement model	Shortest path map based movement
Peers buffer size	50M
Peers' interface	Blue tooth
Message sizes	(500kB - 1MB)
Peers percentage	0.3
Pre-trusted peers weight (<i>init</i>)	0.25

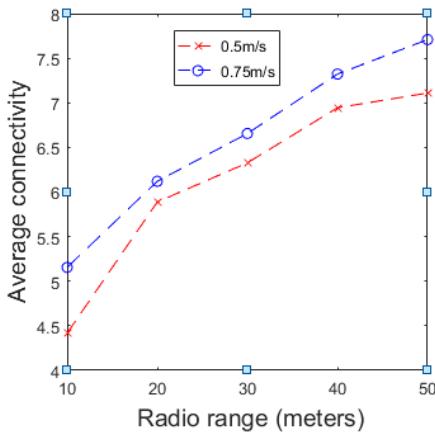


Fig. 3. Average distinct connected neighbours

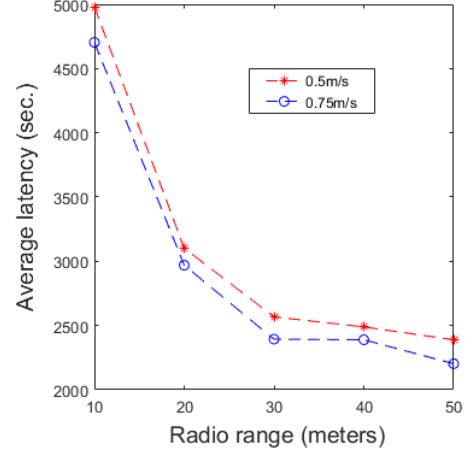


Fig. 5. Average latency in relation to peers' radio ranges

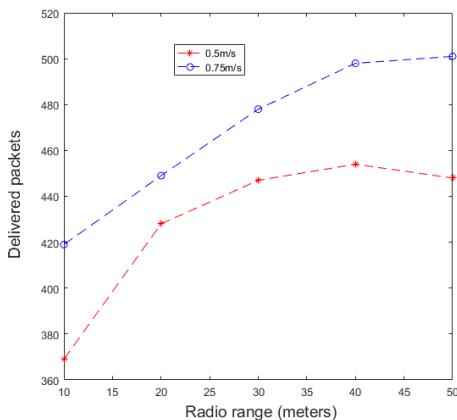
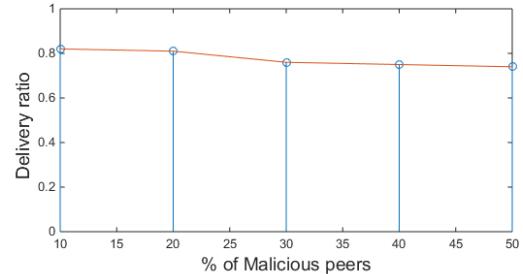


Fig. 4. Delivered packets in relation to peers' radio ranges

higher chances of meeting many other peers and establish connectivity with them. The results further show that, peer-to-peer protocols can exploit the indirect connectivity (transitivity) to reduce the radio ranges in sparse networks thus, reducing the energy required for message transmission. The graph in Figure 4 further shows the number of delivered packets in relation to the increase of peers' interface ranges. The graphs shows that with the increase in peers interface ranges, the number of packets successfully delivered is increasing. In



other words the total number of packets received by all the peers in the network. The higher the number of packets delivered the higher the performance of the protocol. From the presented results shown in Figure 4, one can observe that there is a significant improvement in the number of packets delivered with the increase in the peers' speed. It is important to emphasise here that the number of packets delivered in collaborative networks is closely associated with quality of service considerations, and it is related to reliable network performance.

These results support our arguments that the peers' connectivity is an important factor for efficient trust evaluation between peers in the network. The results further support that, if highly connected peers can determine their corresponding

neighbouring peers with similar a connectivity index, the peers collaborative routing performance can be enhanced. In the next subsection, we proceed to evaluate our proposed protocol based on the peers' connectivity similarity for trust evaluation.

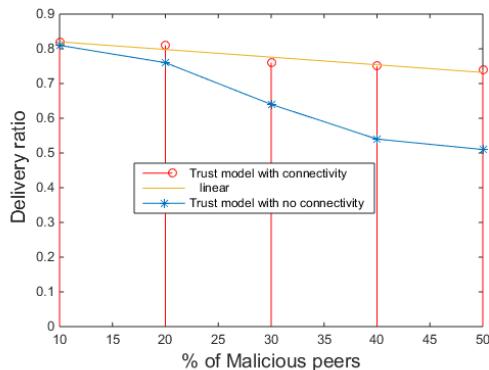


Fig. 7. Benefit of Trust Based Peer Selection Strategy

Next, we focus our attention to the trust evaluation used to optimise the routing performance and to diminish the effect of selfish behaving peers. We consider one of the most important performance metrics (delivery ratio) for the secure implementation of D2D routing protocols. We assigned certain percentage of peers to be periodically dropping the received messages. The malicious peers have limited transmission ranges and buffer sizes. This induces the malicious peers to frequently drop the received packets. We also limit the TTL of the packets created by malicious peers to be 2 minutes only, while for the good behaving peers is up to 300 minutes. This is to enable us to model the behaviour of peers serving as malicious peers. Once the prescribed time count(2 minutes) for packets from malicious peers has elapsed, the peers can discard the packets. Therefore, the number of malicious packets is limited. This is a typical denial-of-service attack which degrade the quality of communication between the peers and reduces network performance. This type of attack can occur due to several reasons; peers being compromised by attackers, peers malfunctioning or any selfish behaviour that can warrant a peer to refuse to participate in packet forwarding. Our goal is to find the best way for peers to identify reliable trustworthy peers for message deliverance.

The experiment proceeds by repeatedly increasing the percentage of malicious peers who drop the received packets frequently. From the result of the experiment, fig 6 shows the maximum delivery ratio obtained when the trust algorithm operates under the best trust formation settings identified in table II. We account the delivery ratio as the total number of packets sent by all the peers in the network divided by the total number of packets received by all the peers in the network.

We see that the delivery ratio remains higher even when the percentage of malicious peers keeps increasing. This to some

extent shows the resilience of our proposed connectivity trust model with the increase in malicious peers.

We then proceed to conduct a comparative analysis, contrasting between the trust model with connectivity scaling factor and a trust model with no connectivity scaling factor. From the graph in Figure 7, it can be observed that the packets delivery probability of the trust model with connectivity scaling factor shows a significant improvement with the increase in malicious peers. However, the delivery probability of a trust model with no connectivity scaling factor show the worst performance.

From the graph in Figure 7, we can deduce that, the implementation of our proposed trust model (with connectivity) exhibits higher performance in comparison with a trust model with no connectivity in terms the packets delivery ratio. For instance, the graph shows a linear slight decrease of delivery ratio with respect to the increase in malicious peers. The result also revealed that the inclusion of connectivity as an element of trust evaluation between the peers improve the peers trust evaluation thus, peers can identify the best possible neighbours to interact with.

Moreover, since the delivery probability favours the increase in the similarity of the connectivity, and the fact that the connectivity between the peers determine the delivery performance, this shows that even in the sparse network, our proposed trust model based on connectivity can yield a good performance in determining the best possible peers to collaborate.

VI. CONCLUSIONS AND FUTURE WORK

The paper has presented a trust-based scheme that exploits peers' neighbourhood characteristics to achieve secure and efficient forwarding strategy among peers in D2D HetNETs. Our proposed solution combines the peers' trustworthiness and similarity of peers' neighbourhood coefficient for trust evaluation. Our trust-based protocol design allows the peers to identify the best possible peer to interact in the midst of the peers while moving to maximise the packets delivery and minimise latency. The result of this study, backed by the simulation validation, demonstrated that there is a correlation between peers' connectivity, peers' interface ranges, and peers' speed and those factors can be used for modelling peers' routing behaviour. We understand that as the peers' radio interface increase and peers' speed increases, the peers tend to establish connectivity and transfer messages with less latency and the routing performance keep increasing. Further, the result validation shows that our proposed solution is resilient against malicious peers and achieves higher performance of packet delivery ratio. Although our proposed trust-based routing protocol development is still very much underway, we discern that the preliminary stage presented in this paper may be useful to any ad-hoc networking protocol design. It shows a new way of interpreting peers connectivity and offers insight into how peers' neighbourhood coefficient can be interpreted as an additional scaling factor for trust and reputation protocol design. Based on the presented study in

this paper many questions need further investigation about the peers' characteristics that can improve peers' routing decision. In the next step of this research, we intend to improve our knowledge about how the three parameters: connectivity, peers' interface range and speed can be modelled to understand peers dynamic motion and behaviour for peers' reciprocity and altruism in trust-based routing.

REFERENCES

- [1] Praveen Jayachandran and Matthew Andrews. Minimizing end-to-end delay in wireless networks using a coordinated edf schedule. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.
- [2] Dayong Ye, Minjie Zhang, and Yun Yang. A multi-agent framework for packet routing in wireless sensor networks. *Sensors*, 15(5):10026–10047, 2015.
- [3] Maggie X Cheng, Xuan Gong, Yibo Xu, and Lin Cai. Link activity scheduling for minimum end-to-end latency in multihop wireless sensor networks. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–5. IEEE, 2011.
- [4] Ángel Cuevas, Manuel Urueña, Gustavo De Veciana, Rubén Cuevas, and Noël Crespi. Dynamic data-centric storage for long-term storage in wireless sensor and actor networks. *Wireless networks*, 20(1):141–153, 2014.
- [5] Yan Sun, Hong Luo, and Sajal K Das. A trust-based framework for fault-tolerant data aggregation in wireless multimedia sensor networks. *IEEE Transactions on Dependable and Secure Computing*, 9(6):785–797, 2012.
- [6] Aminu Bello Usman and Jairo Gutierrez. A reliability-based trust model for efficient collaborative routing in wireless networks. In *Proceedings of the 11th International Conference on Queueing Theory and Network Applications, QTNA '16*, pages 15:1–15:7, New York, NY, USA, 2016. ACM.
- [7] Jaydip Sen. *Reputation-and trust-based systems for wireless self-organizing networks*. Aurbach Publications, CRC Press, USA, 2010.
- [8] Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.
- [9] Li Xiong and Ling Liu. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. *Knowledge and Data Engineering, IEEE Transactions on*, 16(7):843–857, 2004.
- [10] Xiaomei Zhang. *Efficient and Quality-Aware Data Access in Mobile Opportunistic Networks*. PhD thesis, The Pennsylvania State University, 2016.
- [11] Peiyuan Yuan, Huadong Ma, Xiang-Yang Li, Shaojie Tang, and Xufei Mao. Opportunistic forwarding with partial centrality. *arXiv preprint arXiv:1208.0186*, 2012.
- [12] Wei-jen Hsu, Debojyoti Dutta, and Ahmed Helmy. 1 csi: A paradigm for behavior-oriented profile-cast services in mobile networks.
- [13] Bing Bai, Zhenqian Feng, Baokang Zhao, and Jinshu Su. Benefiting from the community structure in opportunistic forwarding. *Comput. Sci. Inf. Syst.*, 10(2):865–876, 2013.
- [14] Azzedine Boukerche and Amir Darehshoorzadeh. Opportunistic routing in wireless networks: Models, algorithms, and classifications. *ACM Comput. Surv.*, 47(2):22:1–22:36, November 2014.
- [15] Ari Keränen, Jörg Ott, and Teemu Kärkkäinen. The one simulator for dtn protocol evaluation. In *Proceedings of the 2nd international conference on simulation tools and techniques*, page 55. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.
- [16] Aminu Bello, William Liu, Quan Bai, and Ajit Narayanan. Revealing the role of topological transitivity in efficient trust and reputation system in smart metering network. In *Data Science and Data Intensive Systems (DSDIS), 2015 IEEE International Conference on*, pages 337–342. IEEE, 2015.
- [17] Ying Zhu, Bin Xu, Xinghua Shi, and Yu Wang. A survey of social-based routing in delay tolerant networks: positive and negative social effects. *IEEE Communications Surveys & Tutorials*, 15(1):387–401, 2013.
- [18] Eyuphan Bulut and Boleslaw K Szymanski. Friendship based routing in delay tolerant mobile social networks. In *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, pages 1–5. IEEE, 2010.
- [19] Nam P Nguyen, Thang N Dinh, Sindhura Tokala, and My T Thai. Overlapping communities in dynamic networks: their detection and mobile applications. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 85–96. ACM, 2011.
- [20] Alessandro Mei, Giacomo Morabito, Paolo Santi, and Julinda Stefa. Social-aware stateless forwarding in pocket switched networks. In *Infocom, 2011 Proceedings Ieee*, pages 251–255. IEEE, 2011.
- [21] Jérémie Leguay, Timur Friedman, and Vanja Conan. Dtn routing in a mobility pattern space. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 276–283. ACM, 2005.
- [22] Jun-Won Ho, Matthew Wright, and Sajal K Das. Zonetrust: Fast zone-based node compromise detection and revocation in wireless sensor networks using sequential hypothesis testing. *IEEE Transactions on Dependable and Secure Computing*, 9(4):494–511, 2012.
- [23] Avinash Srinivasan, Feng Li, and Jie Wu. A novel cds-based reputation monitoring system for wireless sensor networks. In *2008 The 28th International Conference on Distributed Computing Systems Workshops*, pages 364–369. IEEE, 2008.
- [24] Tiejian Luo, Su Chen, Guandong Xu, and Jia Zhou. *Trust-based collective view prediction*. Springer, 2013.
- [25] Yifei Wei, F Richard Yu, and Mei Song. Distributed optimal relay selection in wireless cooperative networks with finite-state markov channels. *IEEE Transactions on Vehicular Technology*, 59(5):2149–2158, 2010.
- [26] Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: Social-based forwarding in delay tolerant networks, 2008.
- [27] Orhan Dengiz. *Maximizing connectivity and performance in mobile ad hoc networks using mobile agents*. ProQuest, 2007.
- [28] David Krackhardt. The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations*, 16(1):183–210, 1999.
- [29] Aminu Bello Usman, William Liu, Quan Bai, and Ajit Narayanan. Trust of the same: Rethinking trust and reputation management from a structural homophily perspective. *International Journal of Information Security and Privacy (IJISP)*, 9(2):13–30, 2015.
- [30] Qing Ou, Ying-Di Jin, Tao Zhou, Bing-Hong Wang, and Bao-Qun Yin. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Physical Review E*, 75(2):021102, 2007.
- [31] Aminu Bello, William Liu, Quan Bai, and Ajit Narayanan. Exploring the role of structural similarity in securing smart metering infrastructure. In *Data Science and Data Intensive Systems (DSDIS), 2015 IEEE International Conference on*, pages 343–349. IEEE, 2015.
- [32] Behrouz Jedari and Feng Xia. A survey on routing and data dissemination in opportunistic mobile social networks. *arXiv preprint arXiv:1311.0347*, 2013.
- [33] Chiara Boldrini. Design and analysis of context-aware forwarding protocols for opportunistic networks. In *Proceedings of the Second International Workshop on Mobile Opportunistic Networking, MobiOpp '10*, pages 201–202, New York, NY, USA, 2010. ACM.
- [34] Aminu Bello Usman and Jairo Gutierrez. Trust-based analytical models for secure wireless sensor networks. In *Security and Privacy Management, Techniques, and Protocols*, pages 47–65. IGI Global, 2018.
- [35] Aminu Bello Usman and Jairo Gutierrez. Datm: A dynamic attribute trust model for efficient collaborative routing. Springer, 2018.
- [36] Feng Li, Jie Wu, and Anand Srinivasan. Thwarting blackhole attacks in disruption-tolerant networks using encounter tickets. In *INFOCOM 2009, IEEE*, pages 2428–2436. IEEE, 2009.
- [37] Ari Keränen, Teemu Kärkkäinen, and Jörg Ott. Simulating mobility and dtns with the one. *Journal of Communications*, 5(2):92–105, 2010.
- [38] Jouni Karvo and Jörg Ott. Time scales and delay-tolerant routing protocols. In *CHANTS '08: Proceedings of the third ACM workshop on Challenged networks*, pages 33–40, New York, NY, USA, 2008. ACM.
- [39] Philo Juang, Hidekazu Oki, Yong Wang, Margaret Martonosi, Li Shiuhan Peh, and Daniel Rubenstein. Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet. In *ACM Sigplan Notices*, volume 37, pages 96–107. ACM, 2002.
- [40] Hoang Anh Nguyen, Silvia Giordano, and Alessandro Puiatti. Probabilistic routing protocol for intermittently connected mobile ad hoc network (propicman). In *2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pages 1–6. IEEE, 2007.

Design and Analysis of Low Power Double Tail Comparator for 2-bit Fast ADC

A. Anitha, M.Balaji , Ravishankar Kandasamy & S.K.Ragul Vijayan

Assistant Professor, ECE,

Sri Shanmuga College of Engineering and Technology

Erode, Tamil Nadu

Abstract— Comparator holds a dominant place in fast ADC circuit for the conversion of analog to digital signal. In this modernized digital world every utilization circuits requires an ADC's with low power to consumption. This in turn reflects in the design of comparators during the design process of the fast ADC circuits, scope is due to the higher number of comparator usage. As the technologies are scaling down, the number of transistor per unit area increases, so that the sub threshold leakage current increases which leads to power consumption in any circuit. This sources the project idea to design a comparator. It is presumed during the design that, it consumes low power in its double tail configuration which when replaces an inverter circuit in latch stage of the double tail comparator by a sleepy inverter. This presumption is validated through the analysis of the simulation results. The power consumption of the designed proposed double tail comparator is 30 μ w when compared to 35 μ w in the conventional type. 2-bit flash ADC circuit is designed and analyzed under two different configurations of the double tail comparator. From the results, it is clear that the power consumption of the ADC circuit designed with proposed sleepy inverter based double tail comparator is observed to be 45mw.

Keywords—sleepy inverter; double tail comparator;analog to digital comparator,leakage current

I. INTRODUCTION

In an open loop configuration, the operational Amplifier finds number of nonlinear applications. e.g. comparators, detectors, limiters and digital interfacing devices namely converter. The comparator is a circuit which compares a signal voltage and the reference voltage. The regenerative comparator consists of the positive feedback and is normally used to provide the high gain and the speed. The dynamic comparator is the regenerative comparator and in which the positive feedback can be obtained by connecting the back to back connected inverter circuit and will act as a latch or the memory element. The Dynamic comparator depends on the clock signal and makes the decision based on whether the applied input signal is higher or lower at the applied clock cycle. It finds more application in the modern world such as analog to digital converter, memory bit line detectors and receivers.

The threshold voltage and gate oxide thickness are scaled down by reducing the channel length of the device. Because of the reduction in threshold voltage the leakage current in the device gets increases exponentially [13]. Theoretically the device will starts conducting when gate source voltage is greater than the threshold voltage ($V_{GS} > V_T$). But it will not be applicable for the device at practical condition. When the transistor is in off state the shorter channel length due to the technology scaling causes the sub threshold current to increase. This current causes the power dissipation during the inactive

mode of the transistor. This leakage current is the limiting factor in the transistor scaling, as it gets increases when the number of transistor increases per unit area. The leakage current for the MOSFET is given by

$$I_{DS} = I_{DS0} e^{\frac{(V_{GS}-V_T)}{nV_T}} \left[\frac{(-V_{DS})}{1 - e^{\frac{(-V_{DS})}{V_T}}} \right]$$

$$V_T = V_{TO} - \eta V_{DS} + \gamma \left[(\phi_s + V_{SB})^{0.5} - (\phi_s)^{0.5} \right]$$

In which the I_{DS0} is current at threshold, V_{TO} is the zero bias threshold voltage, γ is linearised body effect coefficient, η represents the effect of V_{DS} on threshold voltage, n is the sub-threshold swing coefficient, V_T is threshold voltage, V_{SB} source to bulk voltage and V_{GS} is gate source voltage respectively. By increasing the threshold voltage the leakage current decreases. This can be achieved by increasing the source to bulk voltage or by decreasing the gate source voltage, drain source voltage. The literature shows that there are number of comparator available such as Conventional Dynamic Comparator, Conventional Double Tail Dynamic Comparator, and Double Tail Comparator. In which this project focuses on the reduction of power in the double tail comparator.

II. DOUBLE TAIL COMPARATOR

The double tail comparator consists of back to back connected inverter circuit as latch and the preamplifier stage in which the two inputs are given. This circuit operates in two modes one is at reset phase and the other is at decision making phase. The circuit produces the output only at decision making phase and at reset phase the output is at zero values. When clock is zero the circuit works in the reset phase, at which both the tail transistors are OFF and both the fn and fp nodes gets charged by the transistor M9 and M12. These two nodes turn ON the M7 and M8 transistor and this makes both the output nodes outp and outn to zero. When clock is at one condition the circuit operates in the decision making phase. During which both fn and fp nodes discharges depending on the given input signal inp and inn. When input inn is higher, then the node fp discharges faster when compared to fp and this fn node charges the fp node again by turning ON the transistor M11.

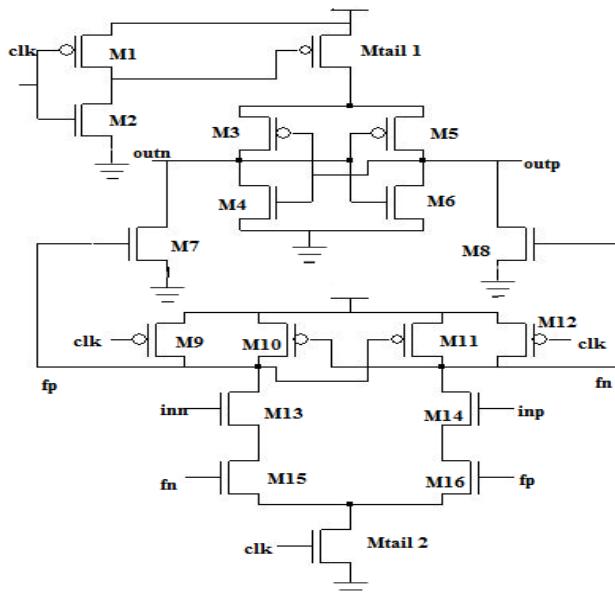


Fig.1 Double Tail Comparator

This makes the outp to zero by switching ON the transistor M8. When this outp is given to the inverter formed by pair of M3 and M4 transistor the outn becomes one. The static power consumption is less in this circuit. The power consumed by this comparator circuit is $35\mu\text{W}$.

III. SLEEPY INVERTER

The inverter circuit in the latch stage of the double tail comparator is replaced by the sleepy inverter circuit. In the sleepy inverter circuit the PMOS is placed below the supply and the NMOS is placed above the ground terminal. This sleepy inverter is operating in two modes such as active mode and sleepy mode.

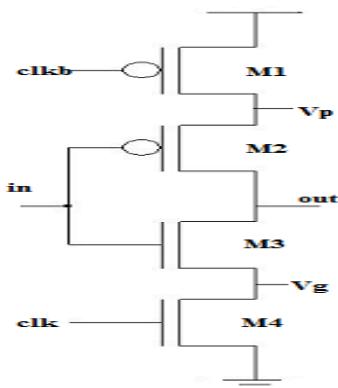


Fig.2 Sleepy Inverter

During the active mode the circuit act as a normal inverter and in sleepy mode there is no current flow in the circuit. When the applied clk is one and clkb is zero, then the circuit operates in the active mode. During which the transistor M1 and M4 conducts and the node Vp is at high potential and the node Vg at ground potential. So the circuit operates as normal inverter. When the applied clk is zero and clkb is one, then the circuit operates in sleepy mode. During which the transistor M1 and M4 turns OFF and the node Vg and Vp is at virtual power

potential and virtual ground potential. The potential at Vg increases and the potential at Vp decreases due to the cut off transistors M1 and M4. The source to body potential of transistor M1 increases, so the threshold voltage increases and hence the leakage current and the power consumption get decreases. The main drawback of the sleepy inverter concept is that, it losses the state information during the sleep mode of operation.

IV. PROPOSED DOUBLE TAIL COMPARATOR

The proposed double Tail Comparator is shown in fig. 3. In which the inverter in the latch circuits is replaced by the sleepy inverter. The operation of the proposed double tail comparator remains same as the double tail comparator.

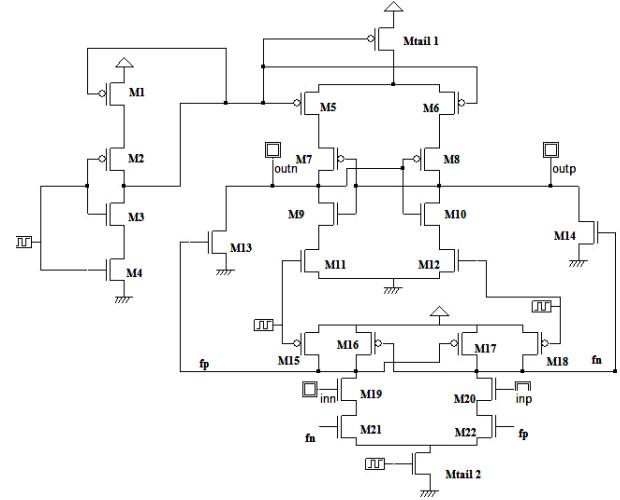


Fig.3 Proposed Double Tail Comparator(with Sleepy inverter)

During normal mode the drain of M5 and M6 is at high potential and drain of M11 and M12 at ground potential. At sleep mode all the M5,M6,M11 and M12 are cut off. By incorporating the sleepy inverter in the proposed double tail comparator the overall power consumption of the circuit is reduced.

V. BLOCK DIAGRAM OF FLASH ADC CIRCUIT

Flash analog to digital converters, also known as parallel ADCs, are the fastest circuit for the conversion of the analog to a digital signal. The Flash type ADC has the least conversion time and is used in time critical applications such as a sample and hold circuit of a digital oscilloscope. The flash type ADC consists of an array of parallel comparators, the potential divider and the priority encoder. In which the analog signal is applied to the non inverting terminal of the comparator and reference voltage is given to the inverting input terminal of the comparator through the potential divider network.

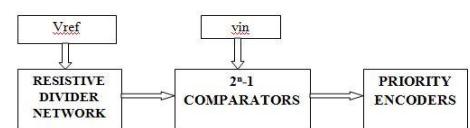


Fig. 4 Block Diagram of Flash ADC circuit

If the n bit digital output is required then $(2^n - 1)$ comparators are used. The priority encoder accepts a 2^n line input and gives out an n bit binary output. Each of the input line has a progressively increasing priority

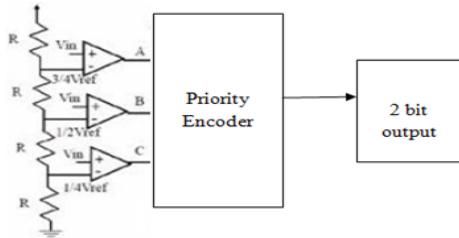


Fig. 5 Block Diagram of Two bit Flash ADC circuit

The two bit flash ADC circuit requires three comparators and priority encoders. This ADC circuit is designed with the designed double tail comparator. When the analog input is between 0 and 1/4 of V_{ref} all comparator produce zero output. When the analog input is between 1/4 and 1/2 of V_{ref} the comparator C only produces 1. When it is between 1/2 and 3/4 of V_{ref} the comparator C and B only produces 1. However, as comparator for C goes high, all comparator below C go high as well. So, a priority encoder is used to convert these input lines from the comparator into binary coded output. The priority encoder includes the priority function. The operation of the priority encoder is such that if two or more inputs are equal to 1 at the same time, the input having the highest priority will take the precedence.

VI. RESULT AND COMPARISON

Both the double tail comparators are simulated with Tanner tool of V14.1. The two comparator circuit is applied with different dc input voltage of inn and inp . In the flash ADC circuit the comparator is fed with the analog signal of amplitude of 30v and the reference voltage of 10v.

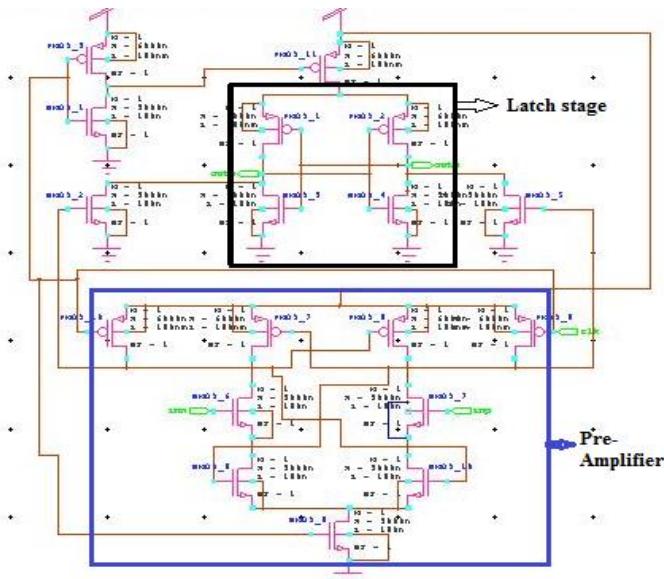


Fig.6 Simulated Double tail comparator

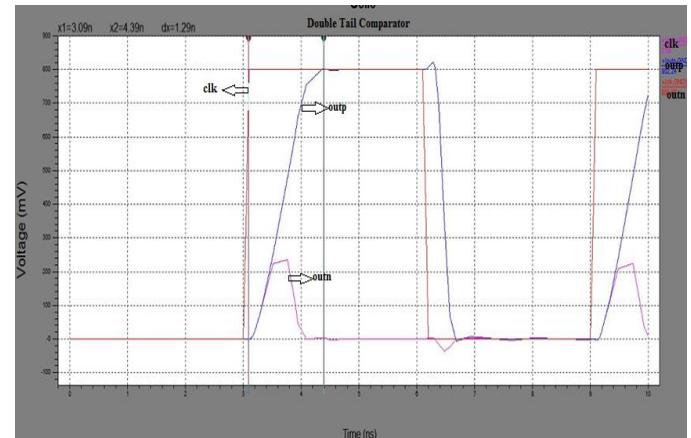


Fig. 7 Output Waveform of Double Tail Comparator

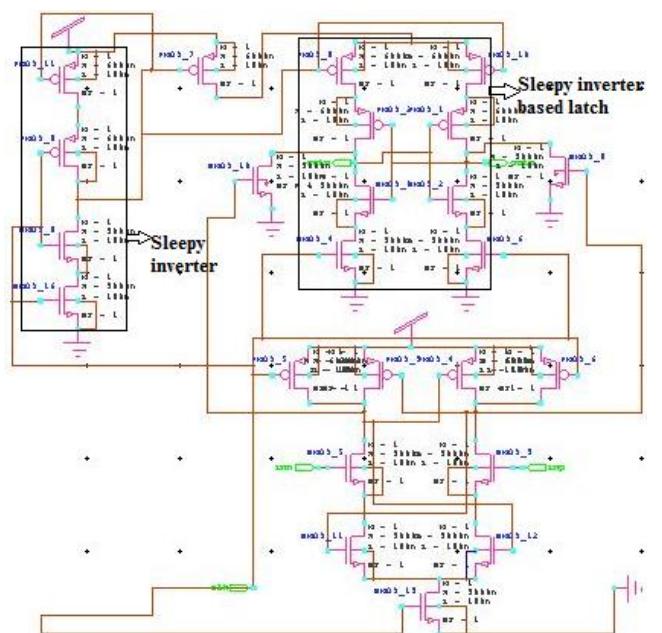


Fig. 8 Simulated Proposed Double Tail comparator

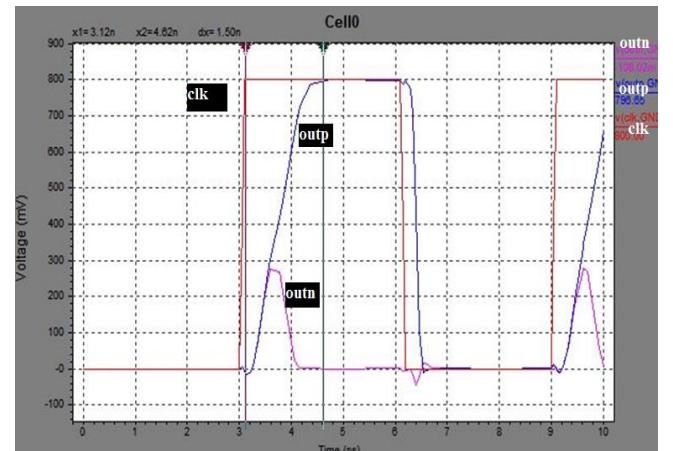


Fig.9 Output waveform of Proposed double tail comparator

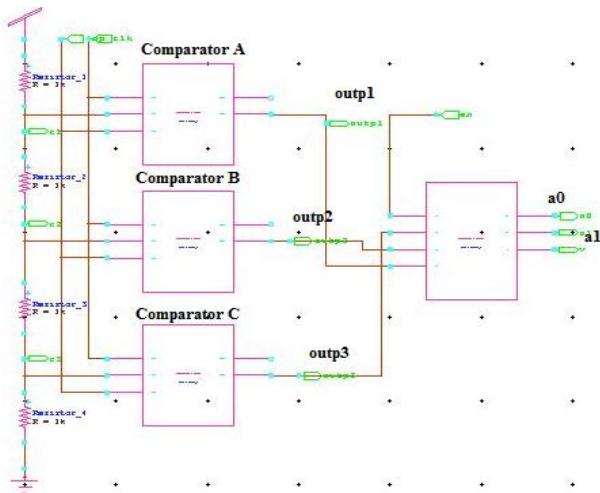


Fig.10 Designed Two bit Flash ADC Circuit

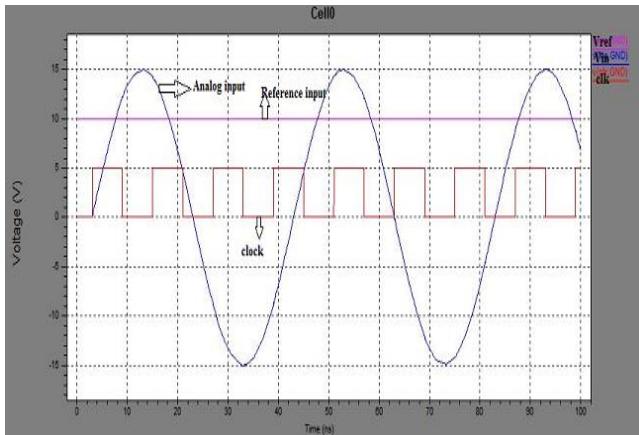


Fig.11 Input Waveform of two bit Flash ADC circuit

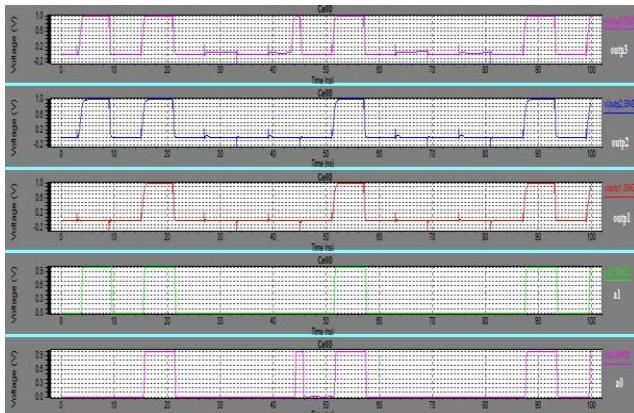


Fig.12 Output and intermediate node Waveform of the Two Bit Flash ADC circuit

The comparison chart shows the power consumed and the delay obtained by the double tail and the proposed double tail comparator. And the power consumed by the different ADC circuit.

Table I

Comparison of Two Designed Comparator

PARAMETERS	MDTDC	PDTDC
Technology(nm)	180	180
Power(μ w)	35	30
Delay(ns)	1.29	1.50
Transistor Count	18	24
Supply Voltage(v)	0.8	0.8

Table II

Comparison of Two Bit Flash ADC with Different Comparators

Circuit	Power consumption	No of transistor
ADC with Modified Double Tail Dynamic Comparator	4.6×10^{-4}	92
ADC with proposed Comparator	4.5×10^{-4}	110

VII. CONCLUSION

All the comparator and the two bit Flash ADC circuit are designed in 180nm CMOS technology using the tanner tool. The proposed double tail comparator designed using the sleepy inverter circuit has the power of 30 μ w and the two bit flash ADC circuit that is designed with this comparator consumes the power of 45mw which is less when compared to the ADC circuit designed with double tail comparator. So the proposed double tail comparator is best suitable for the two bit flash ADC circuit.

REFERENCES

- [1] Abhishek Rai and B. Ananda Venkatesan(2014), "Analysis and Design of High Speed Low Power Comparator in ADC," IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939.

- [2] Abhishek Singh, Meenu Singh, Ajay Dagar and Ashish Mishra (2014), "Optimization of Comparator for High Speed Flash ADC," International Journal of Advanced Technology in Engineering and Science Volume No.02, Issue No. 06.
- [3] P. Arunkumar, G. Chavan, Rekha and P. Narashimaraja(2012), "Design of a 1.5-V, 4-bit Flash ADC using 90nm Technology," International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-2.
- [4] B. Goll and H. Zimmermann(2009) "A comparator with reduced delay time in 65-nm CMOS for supply voltages down to 0.65," IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 56, no. 11, pp. 810–814, Nov. 2009.
- [5] V. Kowsalya(2014),"Design of A Low Power Double Tail Comparator Using Gated Clock and Power Gating Techniques," International Journal of Review in Electronics & Communication Engineering (IJRECE) Volume 2 - Issue 1.
- [6] B. Prasanthi and P. Pushpalatha (2014), "Design of Low-Voltage and low-Power inverter based Double Tail Comparator," International Journal of Engineering Research and General Science Volume 2, Issue 5 ISSN 2091-2730.
- [7] S. Madhumathi and J. Ramesh Kumar (2014), "Design And Analysis Of Low Power And High Speed Double Tail Comparator," International Journal of Technology Enhancements And Emerging Engineering Research, Vol 2, Issue 5 76 Issn 2347-4289.
- [8] Mayank Nema and Rachna Thakur (2012),"Design of Low-Offset Voltage Dynamic Latched Comparator," IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 585-590.
- [9] Neil H. E. Weste and David Money Harris (2011),"CMOS VLSI Design".
- [10] Pedro M. Figueiredo and Joao C. Vital (2006), "Kickback Noise Reduction Techniques for CMOS Latched Comparators," IEEE Transactions on Circuits And Systems-II: Express Briefs, Vol. 53, NO. 7.
- [11] Samaneh Babayan Mashhadi and Reza Lotfi (2014), "Analysis and Design of a Low-Voltage, Low-Power Double-Tail Comparator," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 22, No. 2.
- [12] Sunil Jacob and Deepak Joseph Babu, (2014), "Design of a Low Voltage low Power Double tail comparator in 180nm CMOS Technology," American Journal of Engineering Research (AJER) e-ISSN : 2320-0847 p-ISSN : 2320-0936 Volume-3, Issue-9, pp-15-19.
- [13] International Technology Roadmap for Semiconductors, <http://public.itrs.net>
- [14] Rajani H.P and Srimannarayana Kulkarni (2012) 'Novel sleep transistor techniques for Low leakage power peripheral circuits', International Journal of VLSI design & Communication Systems (VLSICS) Vol.3, No.4.
- [15] M. Janaki Rami and S. Malarkann, "Leakage power Reduction and Analysis of CMOS Sequential Circuits," International Journal of VLSI Design & communication Systems (VLSICS) Vol.3. No.1. February 2012.
- [16] S.Kim, S. Kosonocky, D. Knebel, and K. Stawiasz, "Experimental measurement of a novel power gating structure with intermediate power saving mode," in Proc.Int. Symp. Low Power Electron . Des., pp.20-25

Detecting Windows Operating System's Ransomware based on Statistical Analysis of Application Programming Interfaces (API) Functions Calls

Abdulgader Almutairi

Assistant Professor, College of Sciences and Arts at Qassim University – Kingdom of Saudi Arabia (KSA)
azmtierie@qu.edu.sa

Abstract-Malicious software, or malware in short, pose a serious threat to and can severely damage computer systems. A prime example of this was a ransomware that widely exploited a number of systems recently. Ransomware is a dangerous malware, which is used for extorting money and ransom from victims, failing which they could lose their data forever. In this paper, we used statistical analysis of Application Programming Interfaces (API) functions calls in order to detect ransomware adequately. First, we imported API functions calls for numerous ransomware and benign software applications samples, and saved them as strings in strings files. Subsequently, the imported API functions calls were counted and tabulated to generate a dataset. Then, we applied Chi-Square, Paired Sample t-test, and Correlation statistical analysis to our generated dataset. Our statistical analysis was able to detect ransomware effectively with almost 95% accuracy. In addition, we determined the relationship between each pair.

I. INTRODUCTION

Nowadays, computer systems simplify and organize the performing of most day to day tasks in the modern life. Unfortunately, these computer systems are threatened by malicious and hostile software applications called malwares.^{[1][2]} “Malware” is for a portmanteau of “malicious software application,” which negatively impacts computer systems.^{[1][3]} It became one of the major cyber threats currently and can perform malicious actions, including espionage, information stealing.

In this era of computers, malwares consist of viruses, worms, and trojans, and recently, ransomware.^[4] A virus is a malicious software that injects itself into a carrier program in order to deliver harmful and undesired functions.^{[2][5][4]} A worm is similar to a virus, except that it propagates itself through computer networks.^[6] A trojan is a malicious software that looks legitimate but it is not, and it causes damage and opens backdoors to attackers. A ransomware is a malicious software that prevents or limits the owners from accessing their data until they pay a ransom.^{[7][8][9]}

Recently, ransomwares were the cause of huge monetary losses including \$5B worldwide in 2017 alone, which is expected to reach \$11.5B by 2019. Therefore, extreme caution should be taken in order to protect computer systems and networks.^[7]

II. LITERATURE REVIEW

In the area of malware detection, numerous research works have been conducted based on Application Programming Interfaces (API) functions calls, but they have not adequately covered ransomware.

A. Related Research Works

A research work in ^[10] analyzed imported API functions calls in order to detect the malware, but it did not cover encrypt and decrypt API functions calls, which are crucial in ransomware. Another research in ^[11] applied C 4.5 decision tree algorithm with n-grams=6 to API system calls to detect anomalies, and therefore to secure virtual machines. The study in ^[11] did not cover ransomware. An additional research work in ^[12] applied various machine learning algorithms like DT, SVM, and Random Forest to differentiate malware from cleanware. The study in ^[12] collected various malware samples, but not ransomware. Similarly, study in ^[13] applied various machine learning techniques to detect malware, but did not show ransomware in the collected analyzed samples. A research work in ^[4] applied DT machine learning algorithm to several inputs that were collected from various sources like System, Disk, Network, Registry, and Browser. Like the past studies, this study did not apply DT machine learning algorithm to several inputs caused by ransomware. A research in ^[14] analyzed malware based on binary file features after converting it into an image file in order to analyze it. The weakness of this study is that an attacker can play with bits of binary file, which yields a change in the image file and therefore it will fail to detect malware. One more research work in ^[15] applied SVM machine learning algorithm to its own generated dataset of Windows API with various values of n-grams, but it did not show ransomware samples among the collected samples. A study in ^[16] applied machine learning algorithms like DT and SVM to text mining of API. It used t-test statistical distribution to test the significant difference between them. The study in ^[16] did not cover encrypt and decrypt API functions calls, which are used widely by ransomware. Another research work in ^[17] used Markov Chain to analyze API in order to detect malware. Like the previous researches, this study did not cover encrypt and decrypt API functions calls, which are used widely by ransomware. A research in ^[18] used Z statistical distribution to analyze Linux KVM system calls in order to detect anomalies. As well, this research did not cover encrypt and decrypt API functions calls, which are used widely by ransomware.

III. GENERATING DATASET FOR WINDOWS OPERATING SYSTEMS RANSOMWARE AND BENIGN SOFTWARE APPLICATIONS BASED ON THEIR API FUNCTIONS CALLS

Firstly, we collected numerous ransomware and benign software applications samples for Windows Operating System in order to perform and execute the statistical analysis. The names of collected ransomware software applications samples are Locky.exe, Petya.exe, Cerber.exe, TeslaCrypt.exe, Vipasana.exe, WannaCry.exe, Mamba.exe, Petrwrap, and WannaCry_Plus. The names of collected benign software applications samples names are calc.exe, cliconfg.exe, clipbrd.exe, explorer.exe, freecell.exe, notepas.exe, taskmgr.exe, and spider.exe. The ransomware software applications samples were collected from theZoo ^[19] and reverse ^[20] websites, whereas the benign software applications samples were collected from Win systems32 directory.

Secondly, all software applications samples—whether ransomware or benign—were exported into strings throughout the following executable command under Linux operating system terminal: **\$ strings softwareName.exe >> softwareName.exe.str** where the softwareName.exe involves all ransomware and benign software application samples that are listed above. After that, we studied and analyzed the exported strings software applications samples

due to Windows Operating System Application Programming Interfaces (API) functions calls according to the following Win API functions calls families:

1. Internet Connection
2. Shell Code
3. Encrypt, Decrypt and Digital Certificate
4. Embedded Resource
5. Devices
6. Process Thread Mutex Service and Event (Task)
7. Privileges and User
8. Files and Directories
9. Strings Manipulation
10. Registry and Memory Manipulation

The studying and analyzing of exported strings software applications samples due to Windows Operating System Application Programming Interfaces (API) functions calls identify and count API functions calls for each software application sample by using the following executable command under Linux operating system terminal: **\$ cat softwareName.exe.str | grep -i "Keyword" | wc -l**, where Keyword is one of the ten API functions calls families that are listed above. Then, the results are presented in Table 1 below.

TABLE 1
WINDOWS OPERATING SYSTEM API FUNCTIONS CALLS FOR RANSOMWARE AND BENIGN SOFTWARE
APPLICATIONS SAMPLES.

Software / Win API Functions Calls	Internet Connection	Shell Code	Encrypt, Decrypt, and Digital Certificate	Embedded Resource	Devices	Process, Thread, Mutex, Service, and Event (Task)	Privileges and User	Files and Directories	Strings Manipulation	Registry and Memory Manipulation	Class
1. Locky.exe	0	0	1	0	0	14	5	5	13	19	Ransomware
2. Petya.exe	17	1	30	15	5	88	20	135	122	70	Ransomware
3. Cerber.exe	5	2	14	6	1	42	9	42	40	22	Ransomware
4. TeslaCrypt.exe	6	3	11	1	1	25	3	28	44	7	Ransomware
5. Vipasana.exe	4	2	3	9	1	26	2	47	46	15	Ransomware
6. WannaCry.exe	0	0	8	4	0	11	4	27	10	10	Ransomware
7. Mamba.exe	0	5	65	48	68	292	57	234	186	79	Ransomware
8. calc.exe	0	0	0	0	0	8	3	3	10	4	Benign
9. cliconfg.exe	0	0	0	0	0	4	3	3	0	3	Benign
10. clipbrd.exe	0	0	0	0	2	22	10	33	23	15	Benign
11. explorer.exe	0	1	0	0	3	37	7	20	37	45	Benign
12. freecell.exe	0	0	0	0	1	2	3	2	4	9	Benign
13. notepas.exe	0	0	0	0	1	8	5	17	10	10	Benign
14. taskmgr.exe	0	0	0	0	2	23	6	2	17	13	Benign
15. Petrwrap	0	0	21	13	1	27	4	58	17	6	Ransomware
16. WannaCry_Plus	3	0	11	20	0	36	31	47	20	14	Ransomware
17. spider.exe	0	0	0	0	2	11	4	7	15	11	Benign

IV. STATISTICAL ANALYSIS OF DATASET FOR WINDOWS OPERATING SYSTEMS RANSOMWARE AND BENIGN SOFTWARE APPLICATIONS BASED ON THEIR API FUNCTIONS CALLS

In this research paper, we used Chi-Square statistical distribution to study and analyze the relationships between the ransomware and benign software applications, and Windows operating system API functions calls. The null hypothesis (H_0) and alternative hypothesis (H_1) to be studied and analyzed according to Chi-Square distribution in (1) are as follows:

H₀: There is no relationship between ransomware and benign software applications, and Win API functions calls.

H_1 : There is a relationship between ransomware and benign software applications, and Win API functions calls.

The calculated results of Chi-Square distribution for inputs in Table 1 are recorded in Table 2 below. As results show, only **Encrypt/Decrypt/DigitalCert** Win API functions calls family rejects the null hypothesis (H_0) and accepts alternative hypothesis (H_1), which indicates there is a relationship between ransomware and benign software applications, and Win API functions calls, since P-value is 0.03, which is less than 0.05 ($0.03 \leq 0.05$). This clearly stated that there is a 95% chance of detecting ransomware software applications once the software application uses Encrypt/Decrypt/DigitalCert Win API functions calls. On the contrary, all other Win API functions calls families failed to detect a ransomware software application once the software application uses its corresponding Win API functions calls family in chance 95%.

TABLE 2

No.	Win API Functions Calls Family	Value	df	Asymp. Sig.
1	InternetConnection	6.3	5	0.278
2	ShellCode	4.78	4	0.311
3	Encrypt/Decrypt/DigitalCert	17	8	0.03
4	EmbeddedResource	13.43	8	0.098
5	Devices	6.83	5	0.233
6	Process/Thread/Mutex/Service/Event	14.99	14	0.379
7	Privileged/User	9.31	10	0.503
8	Files/Directories	17	13	0.199
9	Strings	12.32	13	0.502
10	Registry/Memory	12.99	14	0.528

In addition, Paired Sample t-test statistical distribution is used to study and analyze whether the mean difference between two sets of Win API functions calls families is zero or not. The null hypothesis (H_0) and alternative

hypothesis (H_1) to be studied and analyzed according to Paired Sample t-test statistical distribution based on equation in (2) and significance level $\alpha = 0.05$ as follow:

$$H_0: \mu_0 = \mu_1$$

$$H_1: \mu_0 \neq \mu_1$$

Besides that, Correlation (r) is calculated for sets of Win API functions calls families according to equation (3) as follows:

The calculated results of Paired Sample t-test statistical distribution and Correlation for Table 1 inputs are recorded in Table 3 below. As results show, P-value for the pair **EncryptDecryptDigitalCert – ProcessThreadMutexServiceEvent** is 0.03, which is less than 0.05, therefore we accepted the null hypothesis (H_0) and rejected alternative hypothesis (H_1). This indicates that the mean difference between the two sets of Win API functions calls families is zero (means are equal). In the same way, the pair has very strong positive correlation, since $r = 0.94$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. Likewise, P-value for the pair **PrivilegedUser – FilesDirectories** is 0.01, which is less than 0.05, therefore we accepted the null hypothesis (H_0) and rejected alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is zero (means are equal). In addition, the pair has very strong positive correlation, since $r = 0.87$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. Similarly, P-value for the pair **Devices – RegistryMemory** is zero, which is less than 0.05, therefore we accepted the null hypothesis (H_0) and rejected alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is zero (means are equal). In addition, the pair has strong positive correlation, since $r = 0.72$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. Similarly, P-value for the pair **PrivilegedUser – Devices** is 0.02, which is less than 0.05, therefore we accepted the null hypothesis (H_0) and rejected alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is zero (means are equal). In addition, the pair has very strong positive correlation, since $r = 0.86$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. Moreover, P-value for the pair **EmbeddedResource – ShellCode** is 0.05, which is less than or equal 0.05, therefore we accepted the null hypothesis (H_0) and rejected alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is zero (means are equal). In addition, the pair has medium positive correlation, since $r = 0.68$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. Likewise, P-value for the pair **RegistryMemory – Strings** is 0.04, which is less than 0.05, therefore we accepted the null hypothesis (H_0) and rejected alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is zero (means are equal). In addition, the pair has very strong positive correlation, since $r = 0.92$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. In contrast, P-value for the pair

InternetConnection – ShellCode is 0.25, which is greater than 0.05, therefore we rejected the null hypothesis (H_0) and accepted alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is not zero (means are not equal). In addition, the pair has low positive correlation, since $r = 0.24$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair. Similarly, P-value for the pair **FilesDirectories – ProcessThreadMutexServiceEvent** is 0.72, which is greater than 0.05, therefore we rejected the null hypothesis (H_0) and accepted alternative hypothesis (H_1). This indicates that the mean difference between two sets of Win API functions calls families is not zero (means are not equal). In addition, the pair has very strong positive correlation, since $r = 0.94$. This depicts that the increase in one family of one member of the pair will increase the other member of the same pair.

TABLE 3

A PAIRED SAMPLE T-TEST STATISTICAL DISTRIBUTION FOR THE MEAN DIFFERENCE BETWEEN TWO SETS OF WIN API FUNCTIONS CALLS FAMILIES.

Pair	t	df	Sig.	Correlation (r)
EncryptDecryptDigitalCert - ProcessThreadMutexServiceEvent	-2.36	16	0.03	0.94
PrivilegedUser - FilesDirectories	-2.73	16	0.01	0.87
InternetConnection - ShellCode	1.2	16	0.25	0.24
Devices - RegistryMemory	-4.1	16	0	0.72
PrivilegedUser - Devices	2.56	16	0.02	0.86
FilesDirectories - ProcessThreadMutexServiceEvent	0.36	16	0.72	0.94
EmbeddedResource - ShellCode	2.17	16	0.05	0.68
RegistryMemory - Strings	-2.21	16	0.04	0.92

V. CONCLUSION

Ransomware is the most hazardous malicious software (malware); therefore, they should be detected to secure computer systems. In this paper, we used statistical analysis of Application Programming Interfaces (API) functions calls in order to detect ransomware sufficiently. We applied Chi-Square, Paired Sample t-test, and Correlation statistical analysis to our generated dataset, which we produced from various ransomware and benign software applications. Our statistical analysis is able to detect ransomware effectively with an accuracy of 95%, especially when the software application uses Encrypt/Decrypt/DigitalCert Win API functions calls family. In addition, we stated that the mean difference between two sets of Win API functions calls families is zero for all pairs, except InternetConnection – ShellCode and FilesDirectories – ProcessThreadMutexServiceEvent. Similarly, correlation for all pairs is positive, which varies between Low, Medium, or Strong. In future work, we plan to extend this study to apply machine learning algorithms for further improvements.

REFERENCES

- [1] A. JadHAV, D. Vidyarthi, and M. Hemavathy, "Evolution of evasive malwares: A survey," *2016 Int. Conf. Comput. Tech. Inf. Commun. Technol. ICCTICT 2016 - Proc.*, pp. 641–646, 2016.
- [2] H. T. Poon and A. Miri, "Scanning for Viruses on Encrypted Cloud Storage," *Proc. - 13th IEEE Int. Conf. Ubiquitous Intell. Comput. 13th IEEE Int. Conf. Adv. Trust. Comput. 16th IEEE Int. Conf. Scalable Comput. Commun. IEEE Int.*, pp. 954–959, 2017.
- [3] S. Anil and R. Remya, "A hybrid method based on genetic algorithm, self-organised feature map, and support vector machine for better

- [4] network anomaly detection," *Comput. Commun. Netw. Technol. (ICCCNT)*, 2013 Fourth Int. Conf., pp. 1–5, 2013.
- [5] N. Miramirkhani, M. P. Appini, N. Nikiforakis, and M. Polychronakis, "Spotless Sandboxes: Evading Malware Analysis Systems Using Wear-and-Tear Artifacts," *Proc. - IEEE Symp. Secur. Priv.*, pp. 1009–1024, 2017.
- [6] A. A. Shaikh, "Attacks on cloud computing and its countermeasures," *Int. Conf. Signal Process. Commun. Power Embed. Syst. SCOPES 2016 - Proc.*, pp. 748–752, 2017.
- [7] F. C. C. Osorio, H. Qiu, and A. Arrott, "Segmented sandboxing - A novel approach to Malware polymorphism detection," *2015 10th Int. Conf. Malicious Unwanted Software, MALWARE 2015*, pp. 59–68, 2016.
- [8] S. Morgan, "Ransomware Damage Report," 2017. [Online]. Available: <https://cybersecurityventures.com/ransomware-damage-report-2017-part-2/>. [Accessed: 16-Mar-2018].
- [9] J. Z. Kolter and M. A. Maloof, "Learning to detect malicious executables in the wild," *Proc. 2004 ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '04*, vol. 7, p. 470, 2004.
- [10] P. Mishra, E. S. Pilli, V. Varadharajan, and U. Tupakula, "NvCloudIDS: A security architecture to detect intrusions at network and virtualization layer in cloud environment," *2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016*, pp. 56–62, 2016.
- [11] M. Belaoued and S. Mazouzi, "Statistical study of imported APIs by PE type malware," *Proc. - 2014 Int. Conf. Adv. Netw. Distrib. Syst. Appl. INDS 2014*, pp. 82–86, 2014.
- [12] P. Mishra, E. S. Pilli, V. Varadharajan, and U. Tupakula, "Securing virtual machines from anomalies using program-behavior analysis in cloud environment," *Proc. - 18th IEEE Int. Conf. High Perform. Comput. Commun. 14th IEEE Int. Conf. Smart City 2nd IEEE Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2016*, no. Vmi, pp. 991–998, 2017.
- [13] R. Tian, R. Islam, L. Batten, and S. Versteeg, "Differentiating malware from cleanware using behavioural analysis," *Proc. 5th IEEE Int. Conf. Malicious Unwanted Software, Malware 2010*, pp. 23–30, 2010.
- [14] I. Firdausi, C. lim, A. Erwin, and A. S. Nugroho, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection," *2010 Second Int. Conf. Adv. Comput. Control. Telecommun. Technol.*, pp. 201–203, 2010.
- [15] X. Han, J. Sun, W. Qu, and X. Yao, "Distributed malware detection based on binary file features in cloud computing environment," *26th Chinese Control Decis. Conf. (2014 CCDC)*, no. 4, pp. 4083–4088, 2014.
- [16] R. Veeramani and N. Rai, "Windows API based Malware Detection and Framework Analysis," ... *Conf. Networks Cyber Secur.*, vol. 3, no. 3, pp. 1–6, 2012.
- [17] G. G. Sundarkumar, V. Ravi, I. Nwogu, and V. Govindaraju, "Malware detection via API calls, topic models and machine learning," *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2015–Octob, pp. 1212–1217, 2015.
- [18] H. L. Hussein, "Static Analysis Based Behavioral API for Malware Detection using Markov Chain," vol. 5, no. 12, pp. 55–64, 2014.
- [19] S. S. Alariti and S. D. Wolthusen, "Detecting Anomalies In IaaS Environments Through Virtual Machine Host System Call Analysis," *2012 Int. Conf. Internet Technol. Secur. Trans.*, pp. 211–218, 2012.
- [20] Y. Nativ, "theZoo." [Online]. Available: <https://github.com/ytisf/theZoo/tree/master/malwares>. [Accessed: 15-Mar-2018].
- Reverse, "reverse." [Online]. Available: <https://www.reverse.it/>. [Accessed: 15-Mar-2018].

A Review on Congestion Control Approaches for Real-Time Streaming Application in the Internet

Abhay Kumar

Department of CSE, J B Institute of
Engineering and Technology,
Hyderabad, India
abhay25.cse@gmail.com

P. V. S. Srinivas

Sreenidhi Institute of Technology and
Science, Hyderabad, India.
pvssrinivas23@gmail.com

Dr. A Govardhan

Department of CSE, Jawaharlal
Nehru Technological University
Hyderabad, India
govardhan_cse@yahoo.co.in

Abstract— In the support of congestion control over the Internet in providing the assurance of the equality between much diverse traffic is a difficult function. The advent of streaming media has offered users with low-latency media content, with higher congestion on the Internet due to stringent bandwidth and latency requirements. Therefore, it is more and more important to resolve the difficulties of increased packet deliver fail reasoned because of congestion and better quality of service for streaming media. In this paper, we propose a review on the congestion control approaches (CCA) for the real-time streaming applications on the Internet. The role of TCP in network congestion control and the characteristics of the original real-time streaming media are discussed. After that, we discuss issues in the media stream and real-time congestion control. The survey will support the understanding of the current congestion mechanism and continue to enhance the expansion of real-time streaming application services.

Keywords- *Internet, Congestion Control Approaches, Real-Time Streaming, Multimedia*

I. INTRODUCTION

The Internet presents end-to-end finest-attempt for data packet services through utilizing IP protocol exclusive of any unambiguous assurance to the quality of service or deliverance assurance. The presentation and steadiness of the Internet depend on the congestion control approaches (CCA) applied in the "transport layer" protocol. The principal function of transport layer is to utilize internet through "Transmission Control Protocol (TCP)", which consists of a congestion window algorithm that runs on the terminal system. As long as no packet is lost, TCP progressively enhances the transfer rate of the traffic source. This allows traffic sources to establish to what extent the bandwidth is accessible in the network without congestion. Packet deliver fail happens while the network is congested. In response, TCP reduced the sending rate to control congestion. Therefore, TCP adds a multiplication-reduction (AIMD) policy [1] by using addition to cooperatively adjust the sending rate from the suggestion of network congestion in the structure of discarding packets. This approaches permits traffic sources to distinguish congestion and "back off" to accomplish a central delivery rate which will be equal to the congestion point's competence. Therefore, a TCP flow is referred to as a "response" [2] congestion indication (eg, a dropped packet) from the network because congestion control reasons the TCP transmitter to reverse off when congestion is identified.

In these day's internet-based real-time applications such as "VoIP", "video conferencing", and "online games" predominantly use "RTP over UDP" or "UDP" to transmit data. As these protocols do not respond to the congestion measures, the use of these protocols is gaining in popularity, compromising Internet stability [1], [7]. Therefore, in order to construct the real-time application extensively used, it is estimated that a general CCA suitable for real-time multimedia will be deployed. The Internet tends to support the maximum transfer rate of traditionally supported applications, whereas media streaming applications require smoother transfer rates and less jitter. Therefore, in order to securely deploy streaming media applications over the Internet, new congestion control algorithms need to be developed to allow fair interactions with other TCP-oriented applications and maintain the steadiness of the Internet.

In the case of network users competing for scarce network bandwidth, the deliver fail rate is very high during severe congestion in data networks [6]. Internet surveys show a growing demand for bandwidth-intensive applications, resulting in an increase in the delivery failure rate across the entire characteristics of the Internet. It should take appropriate approaches to control network congestion, or the network may be in sustained overload, which may lead to the collapse of the network. Implicit CCA employed in the main transport protocols, TCP used by HTTP and FTP, help improve the robustness of the Internet [1], [5], [9]. Delay-sensitive media applications such as media streams and Webcasts that do not use TCP have disrupted this friendly best-effort network by not responding to network congestion [3], [4]. Many algorithms have been recommending to complement TCP's friendliness for uninterrupted media applications [8], [9], [11], [13], [14].

However, the Internet needs to provide some outline of feedback to data traffic originating from the congested links so that it can regulate its transmission rate depending on the accessible bandwidth, effectively managing end-to-end congestion control [16]. Feedback on congestion can be implicit or explicit. In the case of implicit response, the network's transport layer protocol attempts to maintain high throughput by approximation "service time", "end-to-end delay", and "packet deliver fail". The TCP protocol widely used by the Internet [7], [10] implicitly feeds back lost packets over time and repeatedly. Terminal nodes usually deploy explicit feedback. However, relying on end nodes for implicit

or explicit feedback is not enough to achieve a high throughput of the Internet.

The purpose of this survey is to design effective data routing mechanism to detect congestion and reduce packet delivery fail rate efficiently to advance the competence of Internet throughput, to decrease the packet delivery fail rate and to diminish the network bandwidth requirements. The router's efficient routing mechanism can support the reduction of end-to-end congestion control, which is a major concern for Internet traffic and manages high-bandwidth traffic during congestion. It helps application developers and protocol designers provide the best congestion control for Internet traffic.

The following paper is organized to discuss the TCP in internet congestion control in section-2, Real-time streaming on the Internet in section3, Congestion Control in media streaming mechanism in section-4, investigation of related works in section-5 and the conclusion of the paper in section-6.

II. TCP IN STREAMING CONGESTION CONTROL

During high traffic rates, most TCP methods for congestion control cause a reduction in rate. This results in a huge change in send rate, which is identified by the difference in video quality or a large delay in data buffering before being played, creating a large waiting time for the user. TCP ensures very strict reliability and ordering semantics at the expense of end-to-end delay, which is important for legacy applications such as file transfers but when the data is being played to the user in real time, the receiver is useless. Even responses to dropped packets are retransmitted, resulting in higher network load and significantly lower effective throughput.

TCP is not suitable for emerging applications that stream multimedia content. Media streaming applications are both a data transfer rate and a data transfer rate change. When a media streaming application presents interactive media data to a user in real time, changes in the reception rate at the receiver are visible to the user. These changes are handled by buffering at the receiver, but waiting for the desired buffering results delays the response, uncomfortable viewing, and poor interactivity. Because TCP is purely window-based, it can cause data bursts that get worse due to acknowledged transmissions. This bursty nature makes it very difficult and inefficient to preserve the relative timing of the various data frames received at the receiver [14], [20], [21].

Therefore, TCP with implicit end-to-end congestion control using convective multimedia content is not preferred and a modified protocol with explicit congestion control and the following issues must be used:

- Real-time transmission of media data over the network necessitates high network output because it is simplified to make up for missing data than to make up for the greater delay in reception media data. This is not the case with regular data such as files, where the complete arrival of data at its destination is necessary, so protocols for these static data are not suitable for streaming media.

- In today's computer networks, more and more multimedia applications cannot utilize "Transmission Control Protocol (TCP)" congestion control as a result of "stringent latency" and "jitter" necessities. As the level of congestion enlarges, the excellence degradation of these multimedia functions eventually achieves the level at which users cannot receive content.

The conventional delivery fail-based TCP is not appropriate for real-time traffic because its congestion control continually detects the network's accessible bandwidth, establishing the intervalic rotations throughout which network lineups are foremost packed and then it flows down. These queue fluctuations cause random delay components that vary over time, which increases the circulation time and constructs "delay-sensitive communications" problem. There are two complementary ways to solve this problem: "End-to-End Delay", "Control Over Endpoints", and "Active Queue Management (AQM)" to solve problems in routers [12], [26].

In the following, we provide comments on the end-to-end congestion control algorithm proposed by the related work clustering based on the measures used to derive the congestion and the AQM to control traffic jam queuing setbacks in the network.

A. The exploit of RTT to Infer Congestion

In the field of TCP congestion control, a method of reducing the queuing delay is first designed. Therefore, many real-time business algorithms are based on the past literature studies. The initial congestion control algorithm was particularly applied in Jain's groundbreaking work to include end-to-end delay, dating back to 1989 [35]. Since then, numerous delay-based TCP congestion control modifications comprise are proposed, for example, "TCP FAST" [27] and "TCP Vegas" [36], that utilize "RTT measurements" to conclude congestion. It has been shown that when the RTT is used as a congestion metric, low channel utilization can be achieved when there is reverse traffic or competition with delivery fail-based traffic [19]. It is significant to talk about that in the video conferencing environment, the issue of back traffic is essential because the video stream is sent in both directions.

B. The utilize of Delay-Gradient to Infer Congestion

The initiative of using "RTT gradients" to understand congestion has recently been adopted to prevail over the "latecomer effect" described above. Few instances are "Verus" [23] and "CDG" [28]. The "Verus" [23] is designed specifically for cellular networks and changing the burst link capacity makes congestion control design challenges. The "CDG" [26] aims to offer reasonable subsistence of "loss-based traffic" and "low end-to-end latency". Recently, accurate delay gradient measurements have been demonstrated in data center networks using "NIC hardware time stamping" [29].

C. AQM Algorithms to Reduce the Queuing Delays

The queuing delay can also be minimized by adjusting the network buffer size appropriately [31], [32], [33] or using the AQM algorithm, which controls router buffers through

"reducing packets" or "marking" them as ECNs are employed [34]. Although numerous AQM algorithms have been proposed in the earlier period, their acceptance is blocked due to two major problems: (i) they intend to control the standard queue length as a substitute of the lineup delay, and (ii) the ad-hoc limitations must be configured. These problems, as well as the phenomenon of buffer expansion [30], facilitate the revision of the innovative AQM algorithm, for instance, "CoDel" [26] and "PIE" [24], which do not necessitate constraint modification and unambiguously control the queuing delay rather than the queue length.

III. REAL-TIME STREAMING ON INTERNET

The widespread features of media streaming applications consist of the necessitate for "high bandwidth", "smooth data flow", and "low predictability of end-to-end latency" and "latency differences" [3], [9], [10], [22]. In disparity, the Internet is a most excellent-effort network that does not provide superiority of service assurances, large variations in "network bandwidth" and "host processing speed", the extensive distribution of resources, and high workloads. So, in fact, Internet-based applications have undergone major changes in the available bandwidth, latency, and latency differences. Real-time transmission of media data over the network necessitates high network outcomes because it is simple to make up for lost data than to make up for the greater delay in reception of media data. This is not the case with regular data such as files, where the complete arrival of data at its destination is necessary, so protocols for these static data are not suitable for streaming media. Like "TCP", "UDP" it is also common transport layer protocol that establishes more proprietary protocols over this protocol.

The Internet paradigm for transmitting real-time data such as "audio" and "video" is "Real-time Transport Protocol" (RTP). It is integrated into the application-layer "Real-Time Streaming Protocol (RTSP)" that enables the deliverance of multimedia over the Internet. RTSP supports interoperability between media clients and media servers from different vendors. RTSP uses streaming media to decompose media data into multiple packets and adjusts supported over the accessible bandwidth between the client and the server.

In a typical network, congestion occurs when packets carried by the link exceed the bandwidth competence of the network. It can be described as two categories of congestion difficulties that may occur in the network [16], [17].

- In a "single bottleneck problem", numerous transmitters send their data packets to a destination over a router with an imperfect yield of bandwidth. This congestion knows as "dumbbell-shaped congestion problem". Realizing that congestion is primarily reasoned through the "one-way flow" of numerous transmitters.
- In a common "bottleneck problem", numerous transmitters send packets to several recipients over a series of routers. In the transmitter link, its bandwidth cannot meet the data transmission load requirements. It shows that congestion is a butterfly crowded dilemma. In the mutual "bottleneck problem", various recipients can

correspond with numerous transmitters. This can cause the connection to stream in both directions at the same time.

Previous evaluations [18], [19] argue that single bottlenecks, congestion will lead to higher packet delivery fail rates and share bottlenecks. On the other hand, when the traffic load is constant for a moment, congestion will result in low network throughput for bottleneck sharing issues. However, when the traffic load increases, it will also outcome in a higher packet delivery fail rate. The two congestion problems have a common effect, that buffer overflow. As a result, most CCA focus on resolve this "buffer overflow problem".

The "Real-time Control Protocol (RTCP)" is the object of "RTP", providing management services for streaming applications. The main utility of "RTCP" is to present reaction to the excellence of distribution data. The "RTCP" is used for session control, QoS reporting, and media synchronization. Adaptive multimedia transmitters and receivers are designed so that the protocol for streaming media is simulated. The multimedia receiver sends feedback to the transmitter on network congestion, and the transmitter adjusts its transmission rate accordingly.

IV. CONGESTION CONTROL IN MEDIA STREAMING MECHANISM

This section aims to discuss different congestion prediction and routing models to ensure the highest quality of media delivered under a given network condition. It also emphasizes stochastic transmission schemes as part of the model to reduce jitter and burst loss in media transmissions.

Congestion control in media streaming protocols is designed with rate control to ensure that traffic does not exceed congestion-sensitive TCP traffic, which forms a major part of Internet traffic. A TCP-friendly process has been devised that does not use extra bandwidth but responds to consistent TCP connections based on congestion notifications. Therefore, the object of the invention is to provide a rate adaptation mechanism that does not cause TCP traffic in the background to get in trouble and adjust the transmission rate accordingly to achieve "end-to-end congestion control". The "End-to-end congestion control" depends on the acknowledgment from the receiver so that the sender changes its transmission rate. The congestion control mechanism ensures that a TCP connection using AIMD gets its fair bandwidth allocation when congested.

Modifying the router mechanism includes congestion control at the network layer. If the incoming data rate is higher than the outgoing data rate, the router will not be able to accommodate newly arrived packets because there is no buffer space. In this case, the router must decide to drop the incoming packet. Many strategies are used to make this decision. The simplest and most widely deployed is the tailing algorithm, where each packet arriving at the router is queued until the buffer is full. If there is no space unique, new packets arriving at the router are discarded. Although the tailing algorithm is simple to implement, it shows serious interactions with TCP's congestion control mechanism, resulting in poor performance. Tailing might lead to global synchronization and locking.

Another strategy used in routers is "Active Queue Management", which discards packets before the buffer is full, allowing the source to react to initial congestion. RED is a widely deployed active queue management technology that randomly drops packets depend over the standard queue dimension, where the "queue length" is between the least and highest thresholds. RED is not designed to operate on any particular protocol but rather treats the protocol as a sign of congestion to better enforce the protocol.

V. INVESTIGATION OF RELATED WORKS

The limitations discussed above prompt us to propose ways to improve multimedia streaming by achieving better throughput and lower network load through effectively controlling traffic rates, congestion control and reducing retransmissions.

The "Dynamic adaptation" is a dominant approach, although it needs innovative traffic and CCA to accurately discover and properly utilize the obtainable network bandwidth. The CCA should dynamically establish the share of network bandwidth that adaptive applications be able to utilize moderately under competitive traffic conditions [37], [38]. If these approaches are not sensitive enough to challenging for traffic, probability high multimedia data rates can reason for severe network congestion.

Based on the identified TCP congestion control limits [3], [5], [20], this research work on Multimedia Streams will serve as an enhancement to contribute to the following goals:

- An end-to-end strategy is a classic approach to congestion control. In this model, the TCP sender must detect congestion and take action. Thus, all end-to-end congestion approaches can only rely on implicit congested RTT signals, i.e. packet delivers fail and delay variation. However, in the presence of delivering fail-based algorithms, there are many problems with accurate RTT estimation and lack of link utilization. It will contribute to new improvements in congestion prediction and control routing approaches based on end-to-end CCA. This improves multimedia streaming end-to-end latency and minimizes network overhead.
- Another approach to routing assisted congestion control algorithms will help to overcome the problem of tail-queuing. In the drop tail queue, packets are dropped on the bottleneck router after a queue buffer overflow occurs. In the RED management queue, packets will be discarded earlier based on the output of the discard probability algorithm. If queue usage is low, there is almost no packet deliver fail. Conversely, if the queue begins to populate, the drop probability increases proportionately.
- Finally, it will provide a novel TCP fairness queue mechanism for maintaining fair queue management for each flow that can control flow, and for handling flows that can saturate or delay flow to unacceptable flows. The "Fair queuing mechanisms" [12], [25] distribute the bandwidth of the line fairly and provide relatively small

queue delays for short communications to increase throughput.

G. Carlucci et. al. [1] proposed a new congestion control algorithm for RTC depended on the foremost thought of estimation - using "Kalman filter" in the end-to-end unidirectional stoppage dissimilarity understanding by the packet from the transmitter to the destination. It compares the approximation with a "dynamic threshold" and compel the forceful of the controller located at the receiver, which is intended to maintain a low queuing delay, and a deliver fail-based controller at the transmitter when the loss is found. It works. It is adopted by "Google Chrome" for the congestion control. A large number of experimental evaluations show that this algorithm includes queuing delay while providing fairness of the intro and inter-protocols and full link utilization.

S. D'Aronco et al. [2] proposed in favor of "delay-constrained communication" on the most excellent "packet-switched networks" in a new congestion control algorithm. The algorithm maintains a restricted queuing delay when challenging with erstwhile delay-depend streams, avoiding malnourishment while challenging in favor of delivering fail-based streams. It uses the distinguished value-dependend allocated method as congestion control, however: (1) introduces a recent "non-linear mapping" between "empirical delay" and "value functions", (2) combines delay and loss in the sequence of packets based on packet arrival interval Single-price measurement period.

A. Biernacki [6] studied the traffic commencing "120 client-server pairs" in simulated lab surroundings and multiplexed against a particular network link. It shows that the arrangement of traffic is different from that of the "first generation" and "second generation HTTP video systems" and is not similar to the general Internet traffic construction. The traffic volume acquires shows negative correlation and anti-determination, and its distribution purpose is inclined to the right. In addition, it indicates that the traffic produced by users using the same or similar playout policies is positively related and synchronized (clustering), while the traffic from dissimilar play-out policies demonstrates negative or no correlation.

G. Tian et al. [9] formally studied the trade-off between responsiveness and smoothness in "HTTP Adaptive Streaming (HAS)" by analyzing and experimenting. This shows that client-area buffering of video moment is a superior response indication to conduct video alteration. Then proposed a new video rate control algorithm for stability the video rate smoothing and high bandwidth consumption requirements. It shows that a miniature video rate edge is able to result in a dramatic increase in the smoothness of video rates and buffer sizes. It also proposes HAS invents that exertion with multiple servers and wireless associations. This proves that our HAS design is well-organized and vigorous in a practical network impression.

Most of the existing work [1], [2], [4], [5], [7] usually adopts the method of random early detection to adjust the source rate of the video sender in point of network congestion feedback and maintain the TCP friendliness when the network is congested. The transmission. To overcome this problem, most of these models enable the interaction between Sender-

adaptive transmission and router-first packet filtering to achieve low loss, high-quality multimedia delivery over the Internet.

In order to diminish packet delivery fail reasoned through "bandwidth searching", "video congestion control algorithms" can utilize self-assured fault modification to discard dropped packets. Regrettably, if redundancy is added to the dominant part of the contention flow, additional overhead may cause an increase in packet delivery fail in the network, resulting in additional destruction. However, in this article, it has been established through analysis and simulation that the negative impact of balancing additional overhead can be heightened by means of a paradigm of "DiffServ" discard method and handover inferior priority to "FEC packets". The analysis also shows that the benefits of FEC are maximized when the router buffer is small. This is well related to the latency and jitter requirements of multimedia streams. Transferring video above the Internet is a significant part of numerous multimedia applications. Currently, the Internet lacks QoS support. The heterogeneity of network and terminal systems creates several confronts for the invent of video transmission systems.

In order to offer readers an apparent understanding of this intend gap, we review the improvements and drawbacks of the following methods and solutions.

- 1) Congestion control: There are three CCA: "rate control", "rate adaptive video coding", and "rate shaping". The "Rate control methods" fall into three kinds: "source-based", "receiver-based", and "hybrid". The rate control scheme is able to go after a "model-based approach" or a "Probe-based approach". The "Source-based rate control" is mainly for unicast and be able to be based on a "model-based approach" or a "Probe-based approach". If functional to "multicast", "source-based rate control" be able to simply go after "Probe-based approaches". The "Source-based rate control" requires a further element to implement the rate on the video stream. This element can be "rate-adaptive video coding" or "rate shaping". For illustrations unite "source-based rate control" and "rate-adaptive video coding" be able to be establish in [51], [63]. For instances of "source-based rate control" and "rate shaping" take account in [25]. It is proposed to solve the problem of heterogeneity in multicast video based on receiver and hybrid rate control.

The benefit of "receiver-based control" more than "sender-based control" is that the trouble of alteration moves from the transmitter to the receiver, thereby increasing the flexibility and scalability of the service. The "Receiver-based rate control" may go after a "model-based approach" or a "Probe-based approach". The "Hybrid rate control" unites a few of the most excellent characteristics of "receiver-based" and "transmitter-based control" in conditions of service elasticity and bandwidth effectiveness [39], [40]. Although will be able to merely go after the "Probe-based approach". One benefit of the "model-based approach" to "Probe-based approaches" for video multicasting is that it does not necessitate the

substitute of information between the groups beneath the "Probe-based approach". As a result, it removes the processing of every recipient and the bandwidth convention allied through the substitute of information.

- 2) Error control: It acquires the appearance of "Forward Error Correction (FEC)", "Delay Limit Retransmission", "Error Recovery or Error Hiding". There are three FECs: "channel coding", "source coding based FEC", and "joint source/channel coding". The benefit of all FEC methods above TCP is the decline of video communication delay. FEC based source coding be able to accomplish minor latency than "channel coding", and "joint source/channel coding" can accomplish the best presentation in the sense of rate deformation. The inconveniences of every one FEC methods are increased communication rate, different loss of flexibility characteristics. The "Feedback approaches" be able to utilized to increase FEC flexibility. Dissimilar FEC, which includes reiterating to make progress commencing a fail that may not happen, the "retransmission-based scheme" simply retransmit the failed packet. Therefore, "retransmission-based schemes" adapt to different failed features, ensuing in competent utilize of network sources.

This review concludes that future research efforts based on TCP congestion control limits [3], [5], [20] that have been identified can be developed with new enhancements to support end-to-end classical congestion control that will improve multimedia streaming End-to-end latency minimizes network overhead. At the same time, the problem of discarding tail queues also needs to be solved by a different routing-assisted congestion control algorithm and the new TCP fair queuing mechanism. The fair queuing mechanism [12] equitably distributes the bandwidth of the line to provide a relatively small queue delay for short-time communications to increase throughput.

VI. CONCLUSION

The Internet streaming is nowadays an important application utilized by general internet users. Nevertheless, the most excellent-attempt network is differentiated through active and impulsive changes in accessible bandwidth, which unfavorably affects video superiority. Therefore, it is significant that real-time recognition approaches with varying bandwidths ensure that the video adapts to the available bandwidth and transmits at the highest quality. The traditional view is to rely on end-user applications to deploy CCA to accomplish high network exploitation and a certain extent of traffic equality. In this article, we will discuss a systematic and comprehensive overview of TCP congestion control for TCP live streaming, as well as a study of the Internet streaming media mechanism. It ensures that the impact of congestion can seriously affect the fairness of the danger, and even data routing collapse. The router-based queue management scheme can effectively promote fairness goals and manage network congestion to share network resources fairly.

REFERENCES

- [1] G. Carlucci, L. De Cicco, S. Holmer, S. Mascolo, "Congestion Control for Web Real-Time Communication", IEEE/ACM Transactions on Networking Vol. 25, PP. 2629 - 2642, 2017.
- [2] S. D'Aronco, L. Toni, S. Mena, X. Zhu, P. Frossard, "Improved Utility-Based Congestion Control for Delay-Constrained Communication", IEEE/ACM Transactions on Networking, Vol. 25, PP. 349 - 362, 2017.
- [3] J. Luo, J. Jin, Feng Shan, "Standardization of Low-Latency TCP with Explicit Congestion Notification: A Survey", IEEE Internet Computing, Vol. 21, PP. 48 - 55, 2017.
- [4] R. Lübben, M. Fidler, "Service Curve Estimation-Based Characterization and Evaluation of Closed-Loop Flow Control", IEEE Transactions on Network and Service Management, Vol. 14, PP. 161 - 175, 2017.
- [5] Y. G Zhao, B. Zhang, C. Li, C. Chen, "ON/OFF Traffic Shaping on the Internet: Motivation, Challenges, and Solutions" IEEE Network, Vol. 31, PP. 48 - 57, 2017.
- [6] A. Biernacki, "Analysis of aggregated HTTP-based video traffic" Journal of Communications and Networks, Vol. 18, PP. 826 - 836, 2016.
- [7] L. De Cicco, Gaetano Carlucci, Saverio Mascolo "Congestion Control for WebRTC: Standardization Status and Open Issues", IEEE Communications Standards Magazine, Vol. 1, PP. 22 - 27, 2017.
- [8] Y. Li, H. Liu, W. Yang, D. Hu, X. Wang, Wei Xu, "Predicting Inter-Data-Center Network Traffic Using Elephant Flow and Sublink Information", IEEE Transactions on Network and Service Management, Vol. 13, PP. 782 - 792, 2016.
- [9] G. Tian, Y. Liu, "Towards Agile and Smooth Video Adaptation in HTTP Adaptive Streaming", IEEE/ACM Transactions on Networking, Vol. 24, PP. 2386 - 2399, 2016.
- [10] B. Briscoe, A.a Brunstrom, A. Petlund, D. Hayes, et. al., "Reducing Internet Latency: A Survey of Techniques and Their Merits", IEEE Communications Surveys & Tutorials, Vol. 18, PP. 2149 - 2196, 2016.
- [11] K. Bilal, A. Erbad, "Edge computing for interactive media and video streaming", Second International Conference on Fog and Mobile Edge Computing (FMEC), PP. 68 - 73, 2017.
- [12] G. Abbas, Z. Halim, Z. Haq Abbas, "Fairness-Driven Queue Management: A Survey and Taxonomy", IEEE Communications Surveys & Tutorials, Vol. 18, PP. 324 - 367, 2016.
- [13] A. Javadtalab, M. Semsarzadeh, A. Khanchi, S. Shirmohammadi, A. Yassine, "Continuous One-Way Detection of Available Bandwidth Changes for Video Streaming Over Best-Effort Networks", IEEE Transactions on Instrumentation and Measurement, Vol. 64, PP. 190 - 203, 2015.
- [14] Q. M. Qadir, A. A. Kist, Z. Zhang, "A Novel Traffic Rate Measurement Algorithm for Quality of Experience-Aware Video Admission Control", IEEE Transactions on Multimedia Vol. 17, PP. 711 - 722, 2015.
- [15] D. Li, Mingwei Xu, Ying Liu, Xia Xie, Yong Cui, Jingyi Wang, Guihai Chen "Reliable Multicast in Data Center Networks", IEEE Transactions on Computers, Vol. 63, PP. 2011 - 2024, 2014.
- [16] Z. Li, Xiaoqing Zhu, Joshua Gahm, Rong Pan, Hao Hu, Ali C. Begen, David Oran "Probe and Adapt: Rate Adaptation for HTTP Video Streaming At Scale", IEEE Journal on Selected Areas in Communications, Vol. 32, PP. 719 - 733, 2014.
- [17] S. A. Memon, S. R. Hassan, N. A. Memon, "Evaluation of video streaming performance over the peer-to-peer network", International Conference on Collaboration Technologies and Systems (CTS), PP. 413 - 420, 2014.
- [18] E. Grigorescu, C. Kulatunga, G. Fairhurst, "Evaluation of the impact of packet drops due to AQM over capacity limited paths", 21st IEEE International Conference on Network Protocols (ICNP) PP. 1 - 6, 2013.
- [19] M. Gorius, Y. Shuai, T. Herfet, "Dynamic media streaming under predictable reliability", IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, PP. 1 - 6, 2012.
- [20] P. Yang, L. Xu, "A survey of deployment information of delay-based TCP congestion avoidance algorithm for transmitting multimedia data", IEEE GLOBECOM Workshops (GC Wkshps) PP. 18 - 23, 2011.
- [21] C. Perkins and V. Singh, "Multimedia Congestion Control: Circuit Breakers for Unicast RTP Sessions", RFC 8083, RFC Editor, Mar. 2017.
- [22] B. Briscoe et al., "Reducing Internet latency: A survey of techniques and their merits", IEEE Communications Surveys Tuts., vol. 18, no. 3, pp. 2149-2196, 3rd Quart., 2016.
- [23] Y. Zaki, T. Pötsch, J. Chen, L. Subramanian, and C. Görg, "Adaptive congestion control for unpredictable cellular networks", in Proc. ACM SIGCOMM, vol. 45, no. 5, pp. 509-522, 2015.
- [24] R. Pan et al., "Pie: A lightweight control scheme to address the bufferbloat problem", in Proc. IEEE HPSR, pp. 148-155, 2013.
- [25] J. Jiang, V. Sekar, and H. Zhang, "Improving Fairness, Efficiency, and Stability in HTTP-Based Adaptive Video Streaming with Festive", Proc. ACM CoNEXT '12, pp. 97-108, 2012.
- [26] K. Nichols and V. Jacobson, "Controlling queue delay", Queue ACM, vol. 10, no. 5, pp. 20:20-20:34, May 2012.
- [27] K. Jacobsson, L. L. H. Andrew, A. Tang, S. H. Low, and H. Hjalmarsson, "An improved link model for window flow control and its application to FAST TCP", IEEE Trans. Autom. Control, vol. 54, no. 3, pp. 551-564, Mar. 2009.
- [28] D. A. Hayes and G. Armitage, "Revisiting TCP congestion control using delay gradients", in Proc. 10th Int. Conf. Res. Netw., vol. 2, pp. 328-341, 2011.
- [29] L. De Cicco, G. Carlucci, and S. Mascolo, "Understanding the dynamic behavior of the Google congestion control for RTCWeb", In Proc. Int. Packet Video Workshop (PV), pp. 1-8, 2013.
- [30] A. Mansy, B. V. Steeg, and M. Ammar, "Sabre: A Client-Based Technique for Mitigating the Buffer Bloat Effect of Adaptive Video Flows", Proc. ACM MMSys '13, pp. 214-25, 2013.
- [31] D. Wischik and N. McKeown, "Part I: Buffer sizes for core routers", ACM SIGCOMM Comput. Communications Rev., vol. 35, no. 3, pp. 75-78, Jul. 2005.
- [32] G. Raina, D. Towsley, and D. Wischik, "Part II: Control theory for buffer sizing", ACM SIGCOMM Comput. Communications Rev., vol. 35, no. 3, pp. 79-82, Jul. 2005.
- [33] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Part III: Routers with very small buffers", ACM SIGCOMM Comput. Communications Rev., vol. 35, no. 3, pp. 83-90, Jul. 2005.
- [34] K. Ramakrishnan, S. Floyd, and D. Black, The Addition of Explicit Congestion Notification (ECN) to IP, document RFC 3168, Internet Requests for Comments, Sep. 2001.
- [35] R. Jain, "A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks", ACM SIGCOMM Comput. Communications Rev., vol. 19, no. 5, pp. 56-71, Oct. 1989.
- [36] L. S. Brakmo and L. L. Peterson, "TCP Vegas: End to end congestion avoidance on a global Internet", IEEE J. Sel. Areas Communications, vol. 13, no. 8, pp. 1465-1480, Oct. 1995.
- [37] E. Brosh, S. A. Baset, V. Misra, D. Rubenstein, and H. Schulzrinne, "The delay-friendliness of TCP for real-time traffic", IEEE/ACM Trans. Netw., vol. 18, no. 5, pp. 1478-1491, 2010.
- [38] Z. Li et al., "Probe and Adapt Rate Adaptation for HTTP Video Streaming at Scale", IEEE JSAC, vol. 32, pp. 719-33, 2014.
- [39] M. Ghobadi et al., "Trickle: Rate limiting YouTube Video Streaming", Proc. USENIX ATC '12, pp. 191-96, 2012.
- [40] S. Akhshabiet al., "What Happens When HTTP Adaptive Streaming Players Compete for Bandwidth?", Proc. ACM NOSSDAV '12, pp. 9-14, 2012.

Student Alcohol Consumption Prediction: Data Mining Approach

¹Hind Almayyan, ²Waheeda Almayyan

¹Computer Department, Institute of Secretarial Studies, PAAET, Kuwait
hi.almayyan@paaet.edu.kw

²Computer Information Department, Collage of Business Studies, PAAET, Kuwait
wi.almayyan@paaet.edu.kw

Abstract

Alcohol consumption in higher education institutes is not a new problem; but excessive drinking by underage students is a serious health concern. Excessive drinking among students is associated with a number of life-threatening consequences that include serious injuries; alcohol poisoning; temporary loss of consciousness; academic failure; violence, unplanned pregnancy; sexually transmitted diseases, troubles with authorities, property damage; and vocational and criminal consequences that could jeopardize future job prospects. This article describes a learning technique to improve the efficiency of academic performance in the educational institutions for students who consume alcohol. This move can help in identifying the students who need special advising or counselling to understand the danger of consuming alcohol. This was carried out in two major phases: feature selection which aims at constructing diverse feature selection algorithms such as Gain Ratio attribute evaluation, Correlation based Feature Selection, Symmetrical Uncertainty and Particle Swarm Optimization Algorithms. Afterwards, a subset of features is chosen for the classification phase. Next, several machine-learning classification methods are chosen to estimate the teenager's alcohol addiction possibility. Experimental results demonstrated that the proposed approach could improve the accuracy performance and achieve promising results with a limited number of features.

Keywords

Data mining; Data mining; Classification; Student's performance; Feature selection; Particle swarm optimization; Alcohol consumption prediction.

1. INTRODUCTION

Globally, heavy alcohol drinking is associated with premature death, weaker probability of employment, more absence from work, in addition to lost productivity and lower wages. Moreover, alcohol consumption results in approximately 3.3 million deaths each year [1]. It is the third largest risk factor for alcohol-related hospitalizations, deaths and disability in the world. Approximately one in four children younger than 18 years old in the United States is exposed to alcohol abuse or alcohol dependence in the family [2]. Alcohol consumption has consequences for the health and well-being of those who drink and, by extension the lives of those around them.

The relationship between problematic alcohol consumption and academic performance is a concern for decision makers in education. [3] Alcohol consumption has been negatively associated with poor academic performance, [4] and heavy drinking has been proposed as a probable contributor to student attrition from school. [5]

Traditional methods for monitoring adolescent alcohol consumption are based on surveys, which have many limitations and are difficult to scale. Therefore, several approaches have been investigated using conventional and artificial intelligence techniques in order to evaluate the teenage alcohol consumption. In Crutzen et al. [6] a group of Dutch researchers studied the association between parental reports, teenager perception and parenting practices to identify binge drinkers. They designed a binary classifier using alternating decision trees to establish the effectiveness of the results of exploring nonlinear relationships of data. Montaño et al. [7] proposed an analysis of psychosocial and personality variables about nicotine consumption in teenagers. They applied several classification techniques such as RNA Multi-layer perceptron, radial basis functions and probabilistic networks, decision trees, logistic regression model and discriminant analysis. They discriminated successfully 78.20% of the subjects, which indicates that this approach can be used to predict and prevent similar addictive behavior.

Pang et al. [8] applies a multimodal study to identify alcohol consumption in an audience of minors, specifically the users of the Instagram social network. The analysis is based on facial recognition of selfie photos and exploring the tags assigned to each image with the objective of finding consumption patterns in terms of time, frequency and location. In the same way, they measured the penetration of alcohol brands to establish their influence in the consumption behavior of their followers. Experimental results were satisfactory and compliant with the polls made in the same audience, which can lead to use this approach to other domains of public health.

In Bi et al. [9], a study using two machine learning methods to identify effectively the daily dynamic alcohol consumption and the risk factors associated to it. For this, they proposed a Support Vector Machine (SVM) as classifier to establish a function for stress, state of mind and consumption expectancy, differentiating drinking patterns. After that, a fusion between clustering analysis and feature classification was made to identify consumption patterns based on daily behavior of average intake and detect risk factors associated to each pattern. Zuba et al. [10] proposed machine learning approach that use a feature selection method with 1-norm support vector machines (SVM) to help classify college students between high risk and low risk alcohol drinkers and the risk factors associated to the heavy drinkers. This approach could be used to help to detect early signs of addiction and dependence to alcohol in students.

In this article, we are addressing the prediction of teenager's alcohol addiction by using past school records, demographic, family and other data related to student. This article extends the research conducted by Cortez and Silvain in 2008 [11]. This study seeks to establish the correlation between poor academic performance and the use of alcohols among teenagers. We applied several data mining tools and ends of evaluation shows potential of better results. This article suggests a new classification technique that enhances the student performance prediction using less number of attributes than the ones used in the original research. The aim is getting better prediction results using less parameters in the process.

The article starts the suggested approach is presented in Section 3. Section 4 describes the experiment steps and the involved dataset. Section 5 shows the experiment result. The article concluded with conclusion and further research plan.

2. THE PROPOSED APPROACH

Initially, several machine-learning classification methods, which are considered very robust in solving non-linear problems, are chosen to estimate the class possibility. These methods include feed-forward artificial Neural Network with MLP, Simple Logistic multinomial logistic model, Rotation Forest, Random Forest ensemble learning methods and C4.5 decision tree and Fuzzy Unordered Rule Induction Algorithm (FURIA) classifiers. We carried out extensive experimentation to prove the worth of the proposed approach. We analyze the results of the dataset from each of the perspectives of, Accuracy, ROC and Cohen's kappa coefficient. Feature extraction has played a significant role in many classification systems [12]. On this basis, the focus of this section is on the applied feature selection techniques.

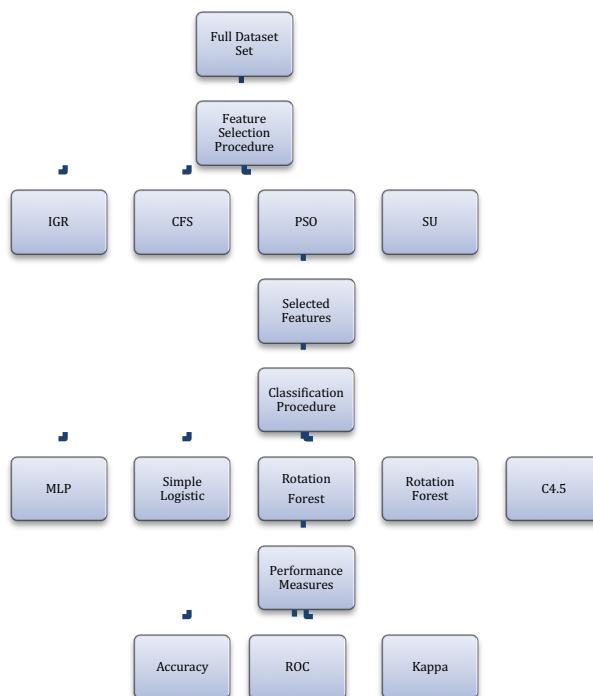


Figure 1 The proposed methodology

2.1 Particle swarm optimization (PSO)

The PSO technique is a population-based stochastic optimization technique first introduced in 1995 by Kennedy and Eberhart [13]. In PSO, a possible candidate solution is encoded as a finite-length string called a particle p_i in the search space. All of the particles make use of its own memory and knowledge gained by the swarm as a whole to find the best solution. With the purpose of discovering the optimal solution, each particle adjusts its searching direction according to two features, its own best previous experience (p_{best}) and the best experience of its companions flying experience (g_{best}). Each particle is moving around the n-dimensional search

space S with objective function $f: S \subseteq \Re^n \rightarrow \Re$. Each particle has a position $x_{i,t}$ (t represents the iteration counter), a fitness function $f(x_{i,t})$ and “flies” through the problem space with a velocity $v_{i,t}$. A new position $z_1 \in S$ is called better than $z_2 \in S$ iff $f(z_1) < f(z_2)$.

Particles evolve simultaneously based on knowledge shared with neighbouring particles; they make use of their own memory and knowledge gained by the swarm as a whole to find the best solution. The best search space position particle i has visited until iteration t is its previous experience p_{best} . To each particle, a subset of all particles is assigned as its neighbourhood. The best previous experience of all neighbours of particle i is called g_{best} . Each particle additionally keeps a fraction of its old velocity. The particle updates its velocity and position with the following equation in continuous PSO [14]:

$$v_{pd}^{new} = \omega * v_{pd}^{old} + C_1 * rand_1() * (pbest_{pd} - x_{pd}^{old}) + C_2 * rand_2() * (gbest_{pd} - x_{pd}^{old}) \quad 1$$

$$x_{pd}^{new} = x_{pd}^{old} + v_{pd}^{new} \quad 2$$

The first part in Equation 1 represents the previous flying velocity of the particle. While the second part represents the “cognition” part, which is the private thinking of the particle itself, where C_1 is the individual factor. The third part of the equation is the “social” part, which represents the collaboration amongst the particles, where C_2 is the societal factor. The acceleration coefficients (C_1) and (C_2) are constants represent the weighting of the stochastic acceleration terms that pull each particle toward the p_{best} and g_{best} positions. Therefore, the adjustment of these acceleration coefficients changes the amount of ‘tension’ in the system. In the original algorithm, the value of $(C_1 + C_2)$ is usually limited to 4 [14]. Particles’ velocities are restricted to a maximum velocity, V_{max} . If V_{max} is too small, particles in this case could become trapped in local optima. In contrast, if V_{max} is too high particles might fly past fine solutions. According to Equation 1, the particle’s new velocity is calculated according to its previous velocity and the distances of its current position from its own best experience and the group’s best experience. Afterwards, the particle flies toward a new position according to Equation 2. The performance of each particle is measured according to a pre-defined fitness function.

2.2 Information Gain Ratio (IGR) attribute evaluation

IGR measure was generally developed by Quinlan [15] within the C4.5 algorithm and based on the Shannon entropy to select the test attribute at each node of the decision tree. It represents how precisely the attributes predict the classes of the test dataset in order to use the ‘best’ attribute as the root of the decision tree.

The expected IGR needed to classify a given sample s from a set of data samples C $IGR(s,C)$ is calculated as follow

$$IGR(s,C) = \frac{gain(s,C)}{split_inf\ o(C)}, \quad 4$$

$$gain(s,C) = entropy(s,C) - entropy_p(s,C),$$

$$entropy(s,C) = -p(s|C) \log_2 p(s|C) - (1-p(s|C)) \log_2 (1-p(s|C)),$$

$$p(s,C) = freq(s|C) / |C|,$$

$$entropy_p(s,C) = \sum_i \frac{|C_i|}{|C|} entropy_p(s,C_i),$$

$$split_inf\ o(C) = -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|},$$

where $freq(s,C)$, C_i and $|C_i|$ are the frequency of the sample s in C , the i^{th} class of C and the number of samples in C_i , respectively.

2.3 Symmetrical Uncertainty

Symmetric uncertainty correlation-based measure (SU) can be used to evaluate the goodness of features by calculating between feature and the target class [16]. The features having greater SU value gets higher importance. SU is defined as

$$SU(X,Y) = \frac{2IG(X|Y)}{(H(X) + H(Y))}, \quad 5$$

$$IG(X|Y) = \frac{H(X)}{H(X|Y)}$$

Where $H(X)$, $H(Y)$, $H(X|Y)$, IG are the entropy of a of X , entropy of a of Y and the entropy of a of posterior probability X given Y and information gain, respectively.

2.4 Genetic Algorithms (GAs)

The basic idea behind the evolutionary algorithms (EAs) is derived from theory of biological evolution developed by Charles Darwin and others. It has been used as computational models and as adaptive search strategies for solving optimization problems. Genetic algorithms were developed in 1975 by Holland as a class of EAs [17]. GAs include a rapidly evolving population of artificial organisms, or so-called agents. Every agent is comprised of a genotype, often called a binary string or chromosome, which encodes a solution to the problem at hand and a phenotype that is the solution. In GAs, at the start the population of agents is randomly generated representing candidate solutions to the problem.

The GAs implementation relies on the appropriate formulation of the fitness function. The main objective of the closed identification fitness function is to maximize the recognition rate. Every agent is evaluated in each iteration, to produce new candidate solutions new fitter offspring and to replace weaker

members of the last generation. Thus, the core of this class of evolutionary algorithms lies in selectively breeding new genetic structures along the course of finding solutions for the problem at hand [18]. We have adopted the algorithm described by Goldberg [19]. The flowchart of GA-based feature selection is described in the Figure 2 below [20].

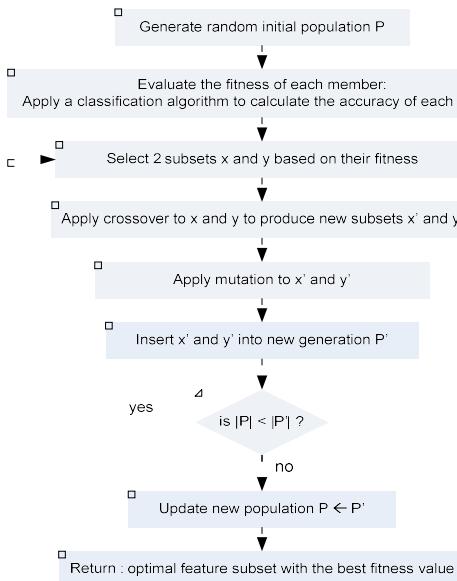


Figure 2 Feature Selection using GA [20]

2.5 Simple random sampling

Usually real-time databases experience class imbalance problems, due to the fact that one class is represented by a considerably larger number of instances than other classes. Subsequently, classification algorithms tend to ignore the minority classes. Simple random sampling has been advised as a good means of increasing the sensitivity of the classifier to the minority class by scaling the class distribution. An empirical study where the authors used twenty datasets from UCI repository has showed quantitatively that classifier accuracy might be increased with a progressive sampling algorithm [21]. Weiss and Provost deployed decision trees to evaluate classification performances with the use of a sampling strategy. Another important study used sampling to scale the class distribution and mainly focus on biomedical datasets [22]. The authors measure the effect of the suggested sampling strategy by the use of nearest neighbor and decision tree classifiers. In Simple random sampling, a sample is randomly selected from the population so that the obtained sample is representative of the population. Therefore, this technique provides an unbiased sample from the original data.

Regarding simple random sampling there are two approaches while making random selection, in the first approach the samples are selected with replacement where the sample can be selected more than once repeatedly with an equal selection chance. In the other approach the selection of samples is done without replacement where the sample can be selected only once, so that each sample in the data set has an equal chance of being selected and once selected it cannot be chosen again [23].

3. Dataset and Evaluation Procedure

3.1 Dataset

The dataset used in this research was collected by customized questionnaire and school reports during the 2005-2006 academic year from two public schools in the Alentejo region of Portugal [11]. The school reports included few attributes such as the three period grades and number of school absences. Researchers have designed a questionnaire with closed questions to extract further socio-demographic information that were expected to affect student performance. Such information includes demographic data (e.g. mother's education, family income), social-emotional (e.g. alcohol consumption) (Pritchard and Wilson 2003) and academic learning attributes (e.g. number of past class failures) that were expected to affect student performance. The questionnaire was first reviewed by school professionals and tested on a small set of 15 students in order to get a feedback. Eventually 788 students completed the customized questionnaire. The dataset has 33 attributes, variables, or features for each student. The academic status or final student performance, which has two possible values: Pass ($G3 \geq 10$) or Fail. Eventually, to find alcohol consumption, there are two different attributes related to alcohol, alcohol taking in work day (D_{alc}) and alcohol taking in weekend(W_{alc}). Therefore, the total alcohol consumption by a specific student in a whole week was estimated using the following formula [24]

$$\text{Alcohol consumption} = (W_{alc} \times 2 + D_{alc} \times 5) / 7$$

6

The new attribute varies between one and five. Therefore, the dataset is divided into two classes according to its alcohol consumption column, which is set to 1 for the alcohol consumption is greater than 3 and 0 otherwise. The 30 features along with description are listed in Table 1.

Table 1: The dataset description of attributes [11]

Attribute Number	Attribute Description	Attribute type	Possible values of attributes
1	School - student's school	Binary	"GP" - Gabriel Pereira or "MS" - Mousinho da Silveira
2	Gender - student's gender	Binary	"F" - female or "M" - male
3	Age - student's age	Numeric	from 15 to 22
4	Address - student's home address type	Binary	"U" - urban or "R" - rural)
5	Famsize - family size	Binary	"LE3" - less or equal to 3 or "GT3" - greater than 3
6	Pstatus - parent's cohabitation status	Binary	"T" - living together or "A" - apart
7	Medu - mother's education	Numeric	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
8	Fedu - father's education	Numeric	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
9	Mjob - mother's job	Nominal	"teacher", "health" care related, civil "services" (e.g. administrative or police), "at _home" or "other"

10	Fjob - father's job	Nominal	"teacher", "health" care related, civil "services" (e.g. administrative or police), "at home" or "other")
11	Reason - reason to choose this school	Nominal	close to "home", school "reputation", "course" preference or "other")
12	Guardian - student's guardian	Nominal	"mother", "father" or "other"
13	Travelttime - home to school travel time	Numeric	1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
14	Studytime - weekly study time	Numeric	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
15	Failures - number of past class failures	Numeric	(numeric: n if 1<=n<3, else 4)
16	Schoolsup - extra educational support	Binary	yes or no
17	Famsup - family educational support	Binary	yes or no
18	Paid - extra paid classes within the course subject	Binary	yes or no
19	Activities - extra-curricular activities	Binary	yes or no
20	Nursery - attended nursery school	Binary	yes or no
21	Higher - wants to take higher education	Binary	yes or no
22	Internet - Internet access at home	Binary	yes or no
23	Romantic - with a romantic relationship	Binary	yes or no
24	Famrel - quality of family relationships	Numeric	from 1 - very bad to 5 – excellent
25	Freetime - free time after school	Numeric	from 1 - very low to 5 - very high
26	Goout - going out with friends	Numeric	from 1 - very low to 5 - very high
27	Health - current health status	Numeric	from 1 - very bad to 5 - very good
28	Absences - number of school absences	Numeric	from 0 to 93
29	G3 - final grade	Numeric	from 0 to 20, output target
30	Alcohol consumption – Target class	Binary	1= yes or 0= no

3.2 Performance Analysis

The performance of the suggested technique was evaluated by using three thresholds and rank performance metrics, Accuracy, ROC and Cohen's kappa coefficient. The main formulations are defined in Equations 6-8, according to the confusion matrix, which is shown in Table 2. In the confusion matrix of a two-class problem, TP is the number of true positives that represent in our case the Pass cases that were classified correctly. FN is the number of false negatives that represents the Pass cases that were classified incorrectly as Fail. TN is the number of true negatives, which represents the Fail cases that were classified as Fail. FP is the number of false positives that represents the Pass cases that were classified as Passed.

Table 2: The confusion matrix

Hypothesis	Predicted patient state	
	Classified Pass	Classified Fail
Hypothesis positive Pass	True Positive TP	False Negative FN
Hypothesis negative Fail	False Positive FP	True Negative TN

Consequently, we can define Precision as:

$$\text{Precision} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\%$$

Precision measures how many of the points predicted as significant are in fact significant. Receiver Operator Characteristic (ROC) curve is another commonly used measure to evaluate two-class decision problems in Machine Learning. The ROC curve is a standard tool for summarizing classifier performance over a range of trade-offs between TP and FP error rates [25]. ROC usually takes values between 0.5 for random drawing and 1.0 for perfect classifier performance.

Kappa error or Cohen's kappa statistics is another recommended measure to compare the performances of different classifiers and henceforth the quality of selected features. Generally, Kappa error value $\in [-1,1]$, so when Kappa error value calculated for classifiers approaches to 1, then the performance of classifier is assumed to be more realistic [26]. The Kappa error measure can be calculated using the following formula:

$$\text{Kappa error} = \frac{P(A) - P(E)}{1 - P(E)} \quad 7$$

where $P(A)$ is total agreement probability and $P(E)$ is the hypothetical probability of chance agreement.

In order to get reliable estimates for classification accuracy on each classification task, every experiment has been performed using 10-fold cross-validation. Cross-validation is a method designed for estimating the generalization error based on "resampling" [27]. Cross-validation technique allows using the whole dataset for training and testing. In k-fold cross-validation procedure, the relevant dataset is partitioned randomly into approximately equal size k parts called folds and trained k times, each time leaving out one of the folds from training process, whilst using only the omitted fold to compute error criterion. Then the average error across all k trials is estimated as the mean error rate and defined as:

$$E = \frac{1}{k} \sum_{i=1}^k e_i \quad 8$$

where, e_i is error rate of each k experiment. Figure 3 depicts the concept behind k-fold cross validation.



Figure 3 Data partitioning using k-fold cross-validation.

The whole dataset is divided into K folds. One-fold ($k=3$, in this example) is set aside to validate the data of testing and the remaining $K-1$ folds are used for training. The entire procedure is repeated for each of the K folds. A number of studies found that the value of 10 for k leads to adequate and accurate classification results [28].

4. Results and discussion

A multi-class classification problem such as predicting student alcohol consumption is a challenging application of data mining. The basic idea of data mining is to extract hidden knowledge using data mining techniques. The suggested system for the purpose of predicting student alcohol consumption applied in this study is carried out in three major phases. The process starts with applying the simple random sampling to scale the imbalanced class distribution. In the second phase, the feature space is searched to reduce the feature numbers and prepare the conditions for the next step. This task is carried out using four feature reduction techniques, namely GR, CFS, SU and PSO Algorithms. At the end of this step a subset of features is chosen for the next round. Afterwards, the selected features are used as the inputs to the classifiers. Five classifiers are proposed to estimate the success possibility as mentioned previously, these methods include MLP, Simple Logistic, Rotation Forest, Random Forest, C4.5 decision tree and SVM.

All the experiments were carried in Waikato Environment for Knowledge Analysis (Weka) a popular suite of data mining algorithms written in Java. The RF algorithm ensemble classifier is designed based on 150 trees and 10 random features to build each tree. While C4.5 classifier was applied with a confidence factor for pruning = 0.25 and a minimum number of instances per leaf of 2. The suggested algorithm is trained using 10-fold cross validation strategy to evaluate the classification accuracy on the dataset. Whereas the PSO feature selection was applied with a population size of 20, number of iterations = 20, individual weight = 0.34 and inertia weight = 0.33.

Table 3 depicts the effect of the class distribution before and after applying the simple random sampling technique. The unbalanced distribution of the two classes makes this dataset suitable to test the effect of simple random sampling strategy. We, therefore, used a simple random sampling approach without replacement to rescale class distribution of the dataset.

The experimental results of the multiple classifiers before and after applying the SRS can be shown in Table 4. The best overall performance is associated with Random Forest classifier with a precision of 92.2%, ROC of 94.5% and Kappa value of 70.4%, and a precision of 98.5%, ROC of 99.7% and Kappa value of 95.2% with all features before and after applying the SRS strategy with all features respectively. As for the classifiers that are used to perform predictions based on the extracted features, we observed that there is no significant difference in performance that explains the importance of SRS step.

Table 3
Class distribution of the Student dataset before and after SRS

Index	Class	Class Distribution	
		Before SRS	After SRS
1	Alcoholic	188	411
2	Not Alcoholic	856	1677

Table 4
Performance measures of selected classifiers before feature selection

Classifier	Performance index	Before SRS	After SRS
MLP	Accuracy	0.846	0.929
	ROC	0.829	0.776
	Kappa	0.479	0.776
Simple Logistic	Accuracy	0.830	0.861
	ROC	0.828	0.874
	Kappa	0.370	0.523
Random Forest	Accuracy	0.922	0.985
	ROC	0.945	0.997
	Kappa	0.702	0.952
C4.5	Accuracy	0.828	0.946
	ROC	0.732	0.933
	Kappa	0.412	0.829
FURIA	Accuracy	0.868	0.967
	ROC	0.824	0.967
	Kappa	0.476	0.885

The second phase involves searching feature vector to reduce the feature numbers and prepare the conditions for the next step. This task is carried out using four feature selection techniques, GR, CFS, SU and PSO Algorithms. The optimal features of these techniques are summarized in Table 5. As noted from Table 3, the dimensionality of features is noticeably reduced. It is worth noting that the number of features has remarkably reduced, therefore less storage space is required for the execution of the classification algorithms. This step helped in reducing the size of dataset to only 6 to 15 attributes.

Table 5
Selected features of the student dataset

FS technique	Number of selected features	Selected features
IGR	13	2,10,11,13,14,15,17,21,25,26,27,28,29
SU	15	2,5,10,11,13,14,15,17,20,21,25,26,27,28,29
GA	7	2,13,25,26,27,28,29
PSO	6	2,13,26,27,28,29

The experimental results of the multiple classifiers with the reduced number of features can be shown in Table 6. The highest performance rate for the IGR-based feature selection technique is associated with Random Forest classifier with 97.9%, 98.7% and 93.3% for Accuracy, ROC and Kappa, respectively with 13 features. While the highest performance rate for the SU-based feature selection technique is associated with C4.5 classifier with 99.5%, 99.7% and 98.2% for Accuracy, ROC and Kappa, respectively with 15 features. The highest performance rate for the GA-based feature selection technique is associated with Random Forest classifier with 96.2%, 97.2% and 88.1% for Accuracy, ROC and Kappa, respectively

with 7 features. The PSO-based feature selection technique highest performance rate is associated with Random Forest classifier with 95.1%, 97% and 84.5% for Accuracy, ROC and Kappa, respectively with 6 features.

Table 6

Performance measures of selected features after SRS

Classifier	Performance index	IGR	SU	GA	PSO
MLP	Accuracy	0.888	0.915	0.861	0.855
	ROC	0.842	0.873	0.839	0.837
	Kappa	0.641	0.716	0.537	0.521
Simple Logistic	Accuracy	0.845	0.857	0.840	0.841
	ROC	0.861	0.874	0.838	0.839
	Kappa	0.476	0.491	0.456	0.461
Random Forest	Accuracy	0.979	0.995	0.962	0.951
	ROC	0.987	0.997	0.972	0.970
	Kappa	0.933	0.982	0.881	0.845
C4.5	Accuracy	0.936	0.984	0.925	0.906
	ROC	0.932	0.986	0.919	0.900
	Kappa	0.797	0.947	0.761	0.698
FURIA	Accuracy	0.962	0.961	0.911	0.890
	ROC	0.970	0.969	0.913	0.848
	Kappa	0.8752	0.8718	0.7046	0.625

Figure 4 visualizes the feature selection agreements between the IGR, SU, GA and PSO models. The Venn diagram shows the suggested models share student's gender, home to school travel time, going out with friends, current health status, number of school absences and final grade, in which all was obtained by the PSO model.

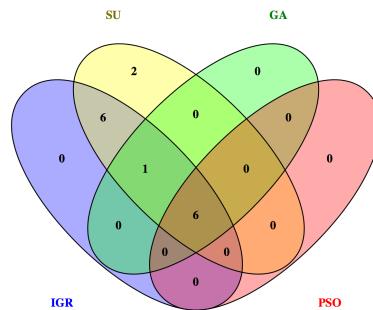


Figure 4 PSO feature selection agreement in the student alcohol consumption dataset

PSO is a well-known optimization method that has a strong search capability and usually used for fine-tuning of the features space. Our proposed technique based on PSO succeeded in significantly improving the classification performance with a limited number of features compared to other techniques. The suggested PSO selection-based features demonstrated accuracies between 84.1% and 95.1% in various DM model and this is a quite high performance for predicting student performance [29]. Therefore, deploying PSO in feature selection clearly helped in reducing the size of dataset from 33 to only 6 attributes. It is worth noting that as the number of features has reduced, less storage space and classification complexity is further required. Moreover, the results demonstrated that these features are adequate to represent the

dataset's class information. The outcomes from the suggested feature selection techniques show better results compared to datasets which are not pre-processed and also when these attribute selection techniques are used independently. As can be seen from above results, the proposed technique based on PSO has produced very promising results on the classification of multi-class dataset in predicting the student alcohol consumption.

As we can comprehend from the data graph and table there is a significant gender differences in the drinking habits. Comparing to men, women are more likely to be responsive to health concerns and are less likely to engage in risky health behaviours [10,11]. Commonly, men smoke and drink more than women in different societies and cultures, and women have a higher expectation of self-control than do men [12,13]. That can lead the other features such as the free time with friends, high travel time between school and home and the number of school absences and eventually the poor academic performance. Our study shows that, drinking is the product of many factors working together. This suggests that the educational professionals can consider these features for further analysis in future.

5. CONCLUSION

Underage drinking or adolescent alcohol misuse is a major public health concern. The proposed machine learning approach could improve the accuracy performance and achieve promising results in identifying risk or protective factors for high-risk drinking that can be used to help detect and address the early developmental signs of alcohol abuse and dependence within adolescent students. The experiment results have shown that the PSO helped in reducing the feature space, whereas adjusting the original data with simple random sampling helped in increasing the region area of the minority class in favour of handling the existing imbalanced data property.

ACKNOWLEDGEMENT

The authors would like to kindly appreciate and gratefully acknowledge, UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>] and Mr. Paulo Cortez [11] for obtaining the student performance dataset.

REFERENCES

1. World Health Organization Management of Substance Abuse Team. Global Status Report on Alcohol and Health. World Health Organization, Geneva, Switzerland; 2011:85.
2. GRANT, B.F. Estimates of U.S. children exposed to alcohol abuse and dependence in the family. American Journal of Public Health 90(1):112–115, 2000.
- 3 Aertgeerts B, Buntinx F. The relation between alcohol abuse or dependence and academic performance in first-year college students. J Adolesc Health. 2002; 31:223–5.
4. Berkowitz AD, Perkins HW. Problem drinking among college students: A review of recent research. J Am Coll Health. 1986;35:21–8.
5. Martinez JA, Sher KJ, Wood PK. Is heavy drinking really associated with attrition from college? The alcohol-attrition

- paradox. *Psychol Addict Behav.* 2008;22:450–6.
6. Crutzen, R., P.J. Giabbani, A. Jander, L. Mercken and H. de Vries, 2015. Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: Building classifiers on adolescent-parent paired data. *BMC Public Health*, 15(1): 747.
7. Montaño, J.J., E. Gervilla, B. Cajal and A. Palmer, 2014. Data mining classification techniques: An application to tobacco consumption in teenagers. *An. Psicol.*, 30(2): 633-641.
8. Pang, R., A. Baretto, H. Kautz and J. Luo, 2015. Monitoring adolescent alcohol use via multimodal analysis in social multimedia. Proceeding of the IEEE International Conference on Big Data (Big Data), pp: 1509-1518.
9. Bi, J., J. Sun, Y. Wu, H. Tennen and S. Armeli, 2013. A machine learning approach to college drinking prediction and risk factor identification. *ACM T. Intell. Syst. Technol.*, 4(4).
10. Zuba, M., J. Gilbert, Y. Wu, J. Bi, H. Tennen and S. Armeli, 2012. 1-norm support vector machine for college drinking risk factor identification. Proceeding of the 2nd ACM SIGHIT International Health Informatics Symposium, pp: 651-660.
11. Cortez, P. and Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In Proceedings of 5th Future Business Technology Conference. pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
12. Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L.A. (2006). Feature Extraction, Foundations and Applications, Springer, Berlin,
13. Kennedy, J. and Eberhart, R. (2001). Swarm intelligence. Morgan Kaufmann.
14. Kennedy, J. and Eberhart, R. (1997). A discrete binary version of the particle swarm algorithm. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol.5, pp. 4104–4108.
15. Quinlan J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.
16. Fayyad, U., and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (pp. 1022–1027). Morgan Kaufmann.
17. J.H. Holland ,Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI (1975).
18. Randy, H., and Haupt, S., 1998. Practical Genetic Algorithms, John Wiley and Sons.
19. David E. Goldberg (1989). Genetic algorithms in search, optimization and machine learning. Addison-Wesley.
20. Hall, Mark A. Correlation-Based Feature Selection for Machine Learning, 1999.
21. G. Weiss and F. Provost, "Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction," *J. Artificial Intelligence Research*, vol.19,315-354,2003.
22. Park, B.-H., Ostrouchov, G., Samatova, N.F., Geist, A.: Reservoir-based random sampling with replacement from data stream. In: SDM 2004 , 492-496, (2004)
23. Mitra SK and Pathak PK. The nature of simple random sampling. *Ann. Statist.*, 1984, 12:1536-1542.
24. Pagnotta, F. and H.M. Amran, 2016. Using data mining to predict secondary school student alcohol consumption. Department of Computer Science, University of Camerino.
25. Fawcett, T. and Provost, F. (1996). Combining data mining and machine learning for effective user profiling. In Proceedings of KDD-96, 8-13. Menlo Park, CA: AAAI Press.
26. Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33: 613–619.
27. Devijver P.A., and Kittler J. (1982). Pattern Recognition: A Statistical Approach. London, GB: Prentice-Hall.
28. Gupta G.K. (2006). Introduction to Data Mining with Case Studies. Prentice-Hall of India.
29. Liao S., Chu, P. and Hsiao, P.(2012). Data mining techniques and applications. A decade review from 2000 to 2011, *Expert Systems with Applications* 39.

Big Data Challenges faced by Organizations

Noureen Kausar

¹Department of Information Technology, GC University
Faisalabad, Pakistan.
nooreallah@hotmail.com

Rabia Saleem

Department of Information Technology, GC University
Faisalabad, Pakistan.
rabia_278@hotmail.com

Sidra Amin

Department of Information Technology, GC University
Faisalabad, Pakistan.
Sidra_chudry44@yahoo.com

Abstract — Big data is a term that describes a large or complex data volume. That data volume can be processes using traditional data processing software or techniques that are insufficient to deal with them. But big data is often noisy, heterogeneous, irrelevant and untrustworthy. As the speed of information growth exceeds Moore's Law at the beginning of this new century, excessive data is making great troubles to human beings. However this data with special attributes can't be managed and processed by the current traditional software system, which become a real problem. In this paper was discussed some big data challenges and problems that are faced by organizations. These challenges may relate heterogeneity, scale, timelines, privacy and human collaboration. Survey method was used as a theoretical solution framework. Survey method consists of a questionnaires report. Questionnaires report consists of all challenges and problems faced by organizations. After knowing the problem and challenges of organizations, a solution was given to organization to solve big data challenges.

Keywords: *Big data, Heterogeneity, Human Collaboration, Organizations Problems, challenges, Security*

I. INTRODUCTION

"Huge records are like teenage sex: absolutely everyone talks about it. No one without a doubt knows a way to do it. Everyone thinks everybody else is doing it. So, all people claim they're doing it, too." The concept of large information has been endemic inside computer science for the reason those earliest days of computing. "Massive information" at the start intended the quantity of facts that could not be processed (successfully) through traditional database techniques and gear. Every time a new garage medium was invented, the amount of records reachable exploded because it is able to be effortlessly accessed. The explosion of statistics has not been followed with the aid of a corresponding new garage medium [1, 21, 22].

We outline "large statistics" as the quantity of facts just past technology's capability to keep, manage and procedure. These imitations are best found by means of a robust analysis of the information itself, express processing needs and the capabilities of the tools (hardware, software, and strategies) used to research it. As with every new trouble, the realization of

how to continue may additionally result in an advice that new tools want to be cast to carry out the new duties. As little as five years in the past, we have been only deliberating tens to loads of gigabytes of storage for our non-public computers. Today, We're wondering in tens to masses of terabytes. As a consequence, big records are a shifting goal. placed some other manner, it's far that quantity of records that is simply past our instant draw close, e.g., we should paintings tough to shop it, get right of entry to it, manage it, and technique it [2,24]. In august 2010, the white residence, OMB, and ostp proclaimed that huge facts are a national mission and precedence at the side of healthcare and national protection (aip, 2010). the country wide technological know-how foundation, the countrywide institutes of fitness, the u.s. geological survey, the departments of defense and power, and the defense superior research initiatives corporation announced a joint r&d initiative in march 2012 with the intention to make investments greater than \$2 hundred million to increase new big records tools and techniques. Its purpose is to enhance our "...know-how of the technologies had to manipulate and mine big amounts of records; observe that understanding to other medical fields "in addition to cope with the countrywide dreams inside the areas of health power protection, education and researcher [3, 27].

A. Big Data has changed the way

Massive statistics has changed the way that we undertake in doing groups, managements and researches. Statistics-in depth technology especially in statistics-in depth computing is coming into the arena those goals to offer the gear that we need to handle the huge records troubles [4, 25, 26] Facts-extensive science is emerging as the fourth clinical paradigm in phrases of the previous specifically empirical technology, theoretical technological know-how and computational technological know-how. Thousand years in the past, scientists describing the herbal phenomenon only primarily based on human empirical evidences, so we call the science at that point as empirical science [5].

B. Relational database management systems

Relational database management systems and computer facts- and visualization-packages frequently have trouble

managing huge data. The paintings might also require "hugely parallel software program walking on tens, loads, or maybe heaps of servers". What counts as "massive statistics" varies depending on the competencies of the users and their tools, and expanding abilities make huge data a shifting target. "For some organizations, dealing with loads of gigabytes of facts for the primary time may additionally cause a need to rethink facts management alternatives. For others, it could take tens or loads of terabytes before facts size turns into a huge attention" [6].

C. Big Data: What's All the Fuss About?

"Each days, we create as a good deal statistics as we did from the dawn of civilization up until 2003" Eric Schmidt, former Google ceo, the belief of massive records is not completely new. In the end, cfos are conversant in managing mounting volumes of information. So why all of the fuss? The volumes in maximum middle financial packages are big but clearly now not inside the nation-states of the terabytes, petabytes, or even zetta bytes being generated by means of the billions of linked gadgets purchasers, companies, and governments use every day round the arena. Large statistics takes 'large' to an entirely new stage, not just in the quantity of records available, but inside the monetary possibilities that information can generate.

But there are different motives that make the troubles of large records one of a kind and urgent. First, the distance among the opportunities afforded with the aid of large records and an organization's capability to take advantage of it's far widening via the second one. As an example, statistics is expected to grow globally with the aid of forty percent in line with yr. however increase in it spending is languishing at simply 5 percent.

2d, companies are being ravaged concurrently by means of the twin demanding situations of rampant financial, regulatory and marketplace alternate and unheard of volatility - all going on at close to 'twitter-speed'. As an end result, there may be excessive hobby in technologies and strategies that can provide a side and shine a mild on market developments fast beforehand of competitors.

Third, stirred by way of way of large information successes reported inside the retail, healthcare and financial services sectors, amongst others, a few marketplace observers do not forget we are at the point of inflection, i.e. that massive records truly is the catalyst for absolutely new boom possibilities, products and services in the personal area, now not to mention price savings and more effective useful resource allocation for government groups [7].

D. Big Data Challenges by Alexandru

Monetary entities and no longer simplest, had advanced over time new and greater complicated strategies that allows them to look marketplace evolution, their function on the market, the efficiency of supplying their services and/or merchandise and so forth. For being able to accomplish that, a large quantity of records is wanted in order to be mined so that could generate treasured insights. Every yr. the facts

transmitted over the net is developing exponentially. By means of the give up of 2016, cisco estimates that the yearly international statistics site visitors will reach 6.6 zettabytes. The task might be not simplest to "accelerate" the net connections, but also to expand software systems with a view to be capable of deal with big data requests in most effective time. To have a higher understanding of what big information means, the table below represents a comparison among conventional statistics and large facts (know-how big records) [8].

TABLE I. Big Data By Alex

Understanding Big Data	
Traditional Data	Big Data
Documents	Photos
Finances	Audio and Video
Stock Records	3D Models
Personnel files	Simulations
	Location data

This situation gives data about the quantity and the sort of huge statistics. It is difficult to paintings with complicated statistics on trendy database structures or on personal computer systems. Generally it takes parallel software program systems and infrastructure that may manage the process of sorting the quantity of statistics that, for instance, meteorologists want to analyze. the request for extra complicated records is getting higher each yr. streaming data in actual-time is turning into a challenge that ought to be triumph over via those corporations that gives such services, as a way to hold their role on the market. Via collecting records in a digital form, corporations take their improvement to a brand new degree. Analyzing virtual data can speed the method of making plans and can also display styles that may be further used so one can improve techniques. Receiving statistics in real-time about consumer needs is useful for seeing market trends and forecasting.

II. BIG DATA CHALLENGES FACED BY ORGANIZATIONS

Big data challenges that is discussed in my research, has been shown in Figure 1.

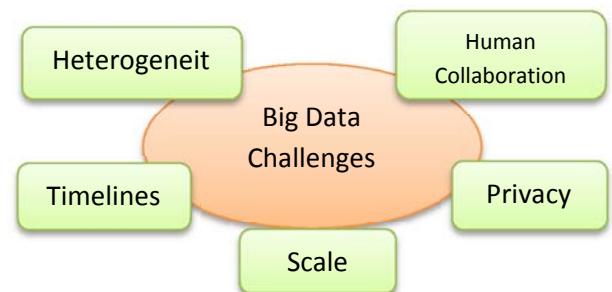


Figure1. Big Data problems

A. Heterogeneity and Incompleteness

Whilst people consume records, a high-quality deal of heterogeneity is without problems tolerated. In truth, the nuance and richness of natural language can offer precious depth. However, gadget analysis algorithms assume homogeneous data, and cannot understand nuance. In consequence, facts have to be carefully established as a primary step in (or previous to) records evaluation. Remember, as an instance, a patient who has multiple medical methods at a medical institution [9].

B. Scale

Of path, the first component everybody thinks of with huge statistics is its length. In spite of everything, the word “big” is there within the very call. Managing large and hastily increasing volumes of facts has been a difficult trouble for many decades. Inside the past, this challenge was mitigated through processors getting quicker, following Moore’s regulation, to offer us with the assets needed to address growing volumes of statistics. But there is an essential shift underway now: information quantity is scaling quicker than compute resources, and CPU speeds are static [10].

C. Timeliness

The turn facet of length is speed. The larger the information set to be processed, the longer it'll take to investigate. The layout of a machine that efficiently deals with length is probable also to result in a device that may system a given size of data set faster [11].

D. Privacy

The privateers of facts are another huge concern, and one that increases within the context of big statistics. For digital fitness facts, there are strict legal guidelines governing what can and can't be executed. For different records, regulations, in particular in the us, are less forceful. However, there may be excellent public fear regarding the beside the point use of private information, mainly via linking of statistics from multiple resources. Dealing with privacy is effectively each a technical and a sociological problem, which should be addressed collectively from each views to comprehend the promise of large information [12, 23].

E. Human Collaboration

No matter the first rate advances made in computational evaluation, there remain many patterns that humans can effortlessly detect but laptop algorithms have a hard time finding. Certainly, catches take advantage of exactly this truth to tell human internet users aside from computer programs. Ideally, analytics for big data will no longer be all computational – rather it will likely be designed explicitly to have a human within the loop. The new sub-field of visible analytics is making an attempt to do that, at the least with admire to the modeling and evaluation phase inside the

pipeline. There's similar fee to human input at all stages of the analysis pipeline [13].

The organizations that are includes for survey of big data challenges faced, these are following

- Government College Lahore
- UVAS
- Punjab University
- UET
- GCUF
- University of Agriculture, Faisalabad
- Faisalabad Institute of Cardiology
- FESCO
- Mobilink and Warid Company
- U Phone Company
- Zong Company
- Wateen Telecom

All these organizations have some related problems, but some educational institutions are still not using big data tools for saving data, only some organization using big data tools for saving data.

These are following related problems faced by organizations during saving data

- Eliminate data entry errors
- Test survey designs
- Change mind
- Try to get better result
- Developed in house
- Space
- Missing data
- Redundancy
- Data collection process
- Human collaboration
- Online verifying
- Information missing
- Incomplete data
- Empty source file
- Saving data
- Security

III. MATERIAL AND METHODS

I was survey of big data challenges in twelve organizations. These are all organizations categorized into four broad area i.s.

- Educational Big Data Challenges
- Big Data Challenges in Telecommunication System
- Big Data Challenges in Hospital
- Big Data Challenges in Electrical Power System

These are all organization has been shown in Figure 2.

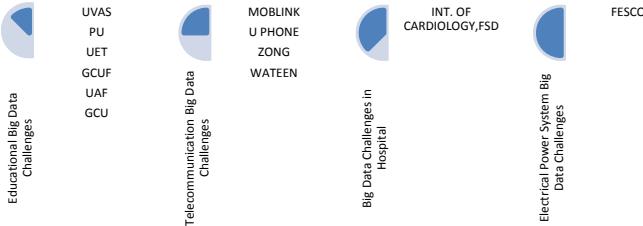


Figure2. Illustrate the survey report about organizations in Pakistan

A. Educational Big Data Challenges

Establishments of better education are running in an increasingly more complex and competitive environment. They may be under increasing pressure to respond to national and global financial, political and social change such as the growing want to boom the percentage of students in sure disciplines, embedding place of business graduate attributes and making sure that the exceptional of gaining knowledge of applications are both nationally and globally applicable [14].

I was survey of many educational institutions of Pakistan. During my survey I realized, educational institutions have faced many problems just because, they are not using Big Data tools for saving their data. All the problems have been shown in Figure 3.

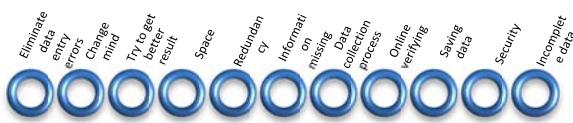


Figure3. Illustrate Big Data related problems in Educational System of Pakistan

B. Big Data Challenges in Telecommunication System

Within the era of Telecommunication, nearly every huge enterprise encounters big information issues, mainly for multinational agencies [15]. On the one hand, the ones corporations commonly have a big variety of customers around the arena. Alternatively, there are very huge volume and speed of their transaction records. For instance, FICO's falcon credit score card fraud detection system manages over 2.1 billion legitimate debts round the arena. There are above three billion portions of content material generated on Facebook every day. The same problem occurs in each internet agencies. The list

may want to pass on and on, as we witness the future agencies warfare fields focusing on large facts [16].

The future of telecommunication has been shown in Figure 4.

#TheFutureOfTelecom

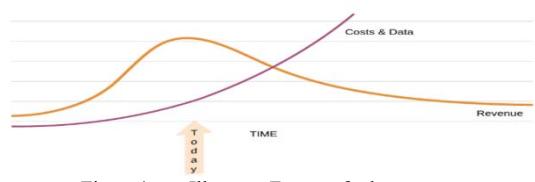


Figure4. Illustrate Future of telecom

C. Big Data Challenges in Hospital

The health network is facing a tsunami of health- and healthcare-associated content generated from several affected person care points of contact, state-of-the-art scientific instruments, and web-primarily based health communities [17, 18].

I was survey of Hospital of Pakistan. During my survey I realized, they have also faced many problems just because, they are not using Big Data tools for saving their data. All the problems have been shown in Figure 5.

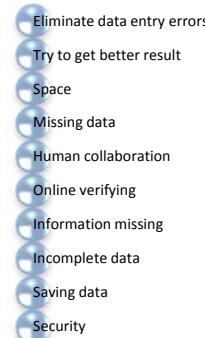


Figure5. Illustrate Big Data related problems in Hospital of Pakistan

D. Big Data Challenges in Electrical Power System

A strength grid is a complicated system connecting an expansion of electrical electricity mills to customers via strength transmission and distribution networks across a massive geographical place, as illustrated in determine 1 [19]. The safety and reliability of energy grids has crucial effect on society and people's each day lifestyles. As an example, on August 14, 2003, a huge part of the Midwest and Northeast United States and Ontario, Canada, skilled an electric powered strength blackout, which affected an area with a population of about 50 million humans. The envisioned overall prices range between \$four billion and \$10 billion (U.S. greenbacks) inside the use, and \$2.3 billion (Canadian dollars) in Canada [20].

I was also survey of FESCO that is an institute of electrical power system of Pakistan. During my survey I realized, they

have also faced many problems just because, they are not using Big Data tools for saving their data. All the problems have been shown in Figure 6.



Figure6. Illustrate Big Data related problems in Electrical Power System

IV. RESULTS AND DISCUSSION

Survey research is one of the maximum essential regions of dimension in carried out social research. The wide location of survey research encompasses any size procedures that involve asking questions of respondents. A "survey" may be something forms a short paper-and-pencil feedback shape to an extensive one-on-one in-depth interview. Survey research is categorized into two broad types' i.s interview and questionnaire. Questionnaire report is consist total 30 questions. I was survey total twelve organizations in Pakistan and filled questionnaire from all these organizations.

Questionnaire were included manage raw data, strategies used for saving data, data saved for future use, Challenges faced during data collection and saving, big data tools are used for saving data, generating source used for data recording, format used for information extracting and changing, method adopt for data cleaning, methods used for querying data, tools used for mining data, interpreted result, type of error you face while managing your data, backup of data, database system used for saving data, power source to use for always on system, type of application, types of database are used for manage your data, type of model, client-server based and type of locking etc.

Organizations were included Moblink, U phone, Zong, Wateen, Cardiology hospital, FSD, GCUF, GCU, PU, UET, UVAS, UAF and FESCO etc.

A. Managing of Raw Data

In this table illustrate that 75% organizations of Pakistan are managed their raw data using computer system, and 25% organizations managed their raw data using both computer and manual system. The results also have been shown in Table 2 and in Figure 7.

TABLE II. Raw Data Management

Manage Raw Data				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Computerized	9	75.0	75.0
	Both	3	25.0	25.0
	Total	12	100.0	100.0

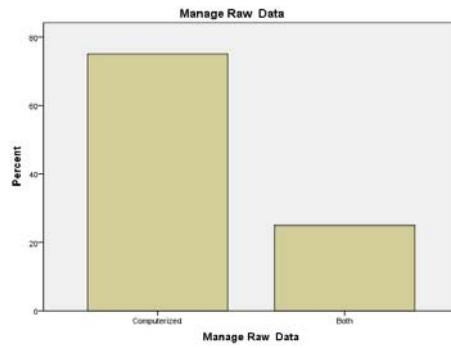


Figure7. Illustrate management of raw data

B. Strategies for Saving Data

In this table shows that 58.3% organizations of Pakistan using backup system, 8.3% using cloud computing, 8.3% using data warehouses, 16.7 using no strategies and 8.3% using others strategies for saving their data. The results also have been shown in Table 3 and in Figure 8.

TABLE III. Strategies For Saving Data

Strategies For Saving Data					
	Frequency	Percent	Valid Percent	Cumulative Percent	
Valid	Backup system	7	58.3	58.3	58.3
	Cloud computing	1	8.3	8.3	66.7
	data warehouse	1	8.3	8.3	75.0
	None	2	16.7	16.7	91.7
	Others	1	8.3	8.3	100.0
	Total	12	100.0	100.0	

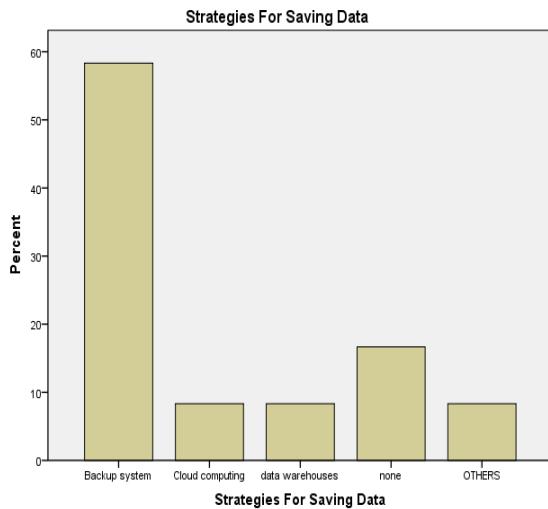


Figure8. Illustrate that strategy for saving data

C. Others Strategies For Saving Data

This table shows that others strategies for saving data. Almost 8.3% organizations of Pakistan using cloud server and Oracle to save their data. The results also have been shown in Table 4 and in Figure 9.

TABLE IV. Others Strategies for saving data

Others Strategies for Saving Data

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10	83.3	83.3	83.3
cloud server	1	8.3	8.3	91.7
Oracle	1	8.3	8.3	100.0
Total	12	100.0	100.0	

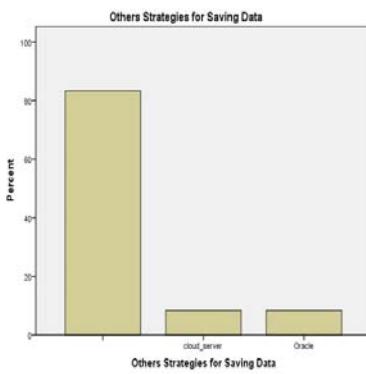


Figure9. Illustrate those others strategies for saving data

D. Big Data Tools For Saving Data

In this table shows that only 16.7% organizations of Pakistan using Hadoop, 8.3 using Jaspersoft and 8.3% using Talend Open Studio and 66.7% organizations not using big data tools for saving data. They are all using others strategies for saving their data. The results also have been shown in Table 5 and in Figure 10.

TABLE V. Big Data Tools For Saving Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hadoop	2	16.7	16.7	16.7
	Jaspersoft BI Suite	1	8.3	8.3	25.0
	Talend Open Studio	1	8.3	8.3	33.3
	Others	8	66.7	66.7	100.0
	Total	12	100.0	100.0	

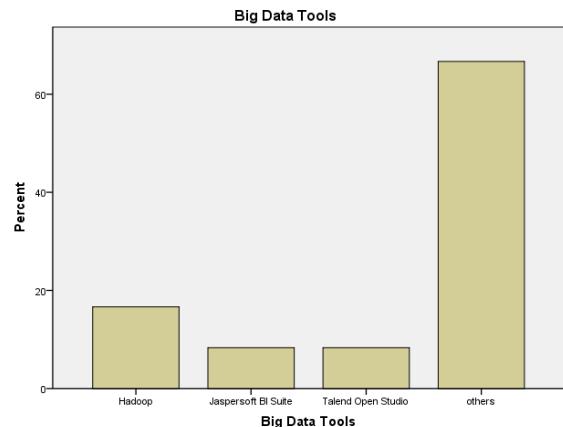


Figure10. Illustrate that big data tools for saving data

E. Using others tools rather than Big Data

In this table shows that 25.0% organizations of Pakistan using cloud computing, 25% using Oracle and 16.7% using SQL for saving their data. The results also have been shown in Table 6 and in Figure 11.

TABLE VI. Others Tools Rather Than Big Data

Using Others tools rather than Big Data

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4	33.3	33.3	33.3
Cloud based	1	8.3	8.3	41.7

Cloud Computing	2	16.7	16.7	58.3
Oracle	3	25.0	25.0	83.3
SQL	2	16.7	16.7	100.0
Total	12	100.0	100.0	

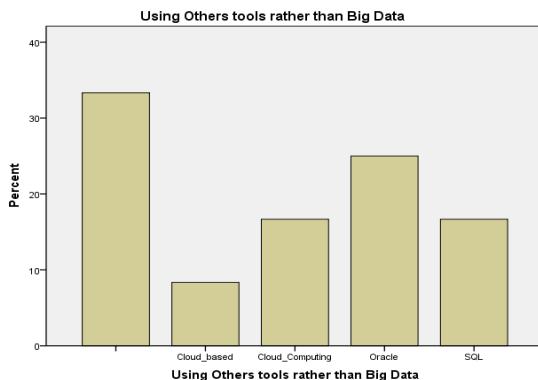


Figure11. Illustrate that organizations using others tools rather than big data

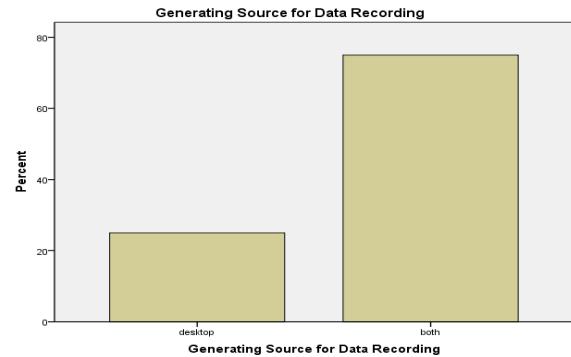


Figure12. Illustrate that data recording sources

G. Format for Information Extracting and Changing

This table shows methods for information extracting and changing, 16.7% organizations of Pakistan using archive file, 8.3% using default compression, 33.3% using data extraction tools and 41.7% using others strategies for information extracting and changing. The results also have been shown in Table 8 and in Figure 13.

TABLE VIII. Information Extracting And Changing

Information Extracting and Changing

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	16.7	16.7	16.7
Archive file				
default compression	1	8.3	8.3	25.0
data extraction tools	4	33.3	33.3	58.3
Others	5	41.7	41.7	100.0
Total	12	100.0	100.0	

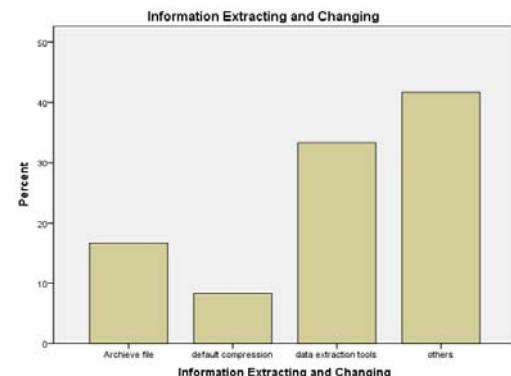


Figure13. Illustrate information extracting and changing

Figure 14 has been shown about power source for system running always. In this figure #1 shows “Generator Power Source”, #2 shows “UPS Power Source”, #3 shows

“Both” and #4 shows “none”. During my survey, I found the result only one educational institution using one source and all others using both source for always running systems. The result has been shown in Figure 14.

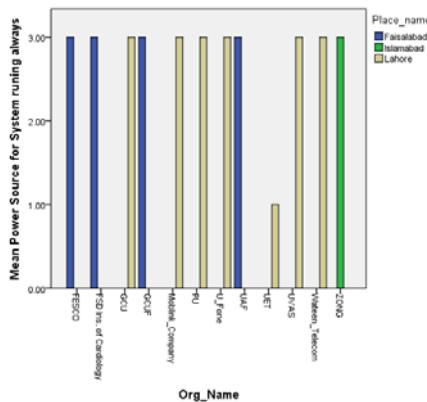


Figure14. Illustrate power source for system running

Figure 15 has been shown about the Database Application System. In this figure #1 shows “web based”, #2 shows “desktop”, #3 shows “manual” and #4 shows “all of these”. During my survey, I found the following result. The result has been shown in Figure 15.

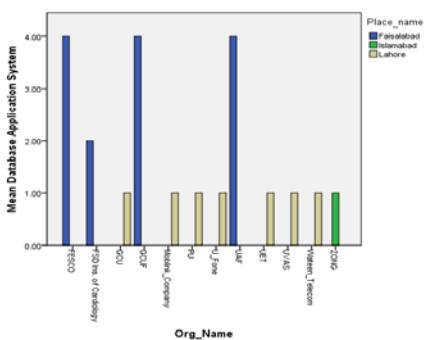


Figure15. Illustrate Database Applications System

Figure 16 has been shown about the Database System. In this figure #1 shows “SQL”, #2 shows “Oracle”, #3 shows “IBM Data Warehouse” and #4 shows “others”. During my survey, I found the following result. The result has been shown in Figure 16.

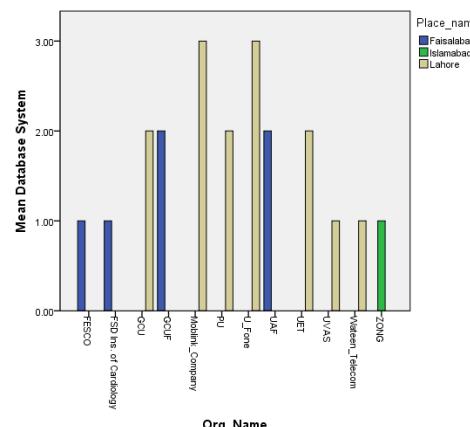


Figure16. Illustrate Database system

Chi test associations between client servers based system and database locking

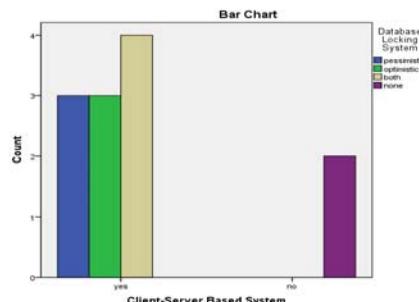
Research Hypothesis (H1) Client-Servers based system and database locking system relate to each other

Significance level=0.05%

			Database Locking System				Total
Client-Server Based System	yes	Count	pessimistic	optimistic	both	none	
		Expected Count	2.5	2.5	3.3	1.7	10.0
	no	% within Client-Server Based System	30.0%	30.0%	40.0%	.0%	100.0%
		Count	0	0	0	2	2
		Expected Count	.5	.5	.7	.3	2.0
	Total	% within Client-Server Based System	.0%	.0%	.0%	100.0%	100.0%
		Count	3	3	4	2	12
		Expected Count	3.0	3.0	4.0	2.0	12.0
% within Client-Server Based System			25.0%	25.0%	33.3%	16.7%	100.0%

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Client-Server Based System * Database Locking System	12	100.0%	0	.0%	12	100.0%



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12.000 ^a	3	.007
Likelihood Ratio	10.813	3	.013
Linear-by-Linear Association	5.124	1	.024
N of Valid Cases	12		

a. 8 cells (100.0%) have expected count less than 5. The minimum expected count is

.33.

Figure 17. Client-Server Based System

*Chi test associations between DBA in Organizations * Experience of DBA*

Research Hypothesis (H1) DBA in Organizations and Experience of DBA
Significance level=0.05%

DBA In Organizations * Experience of DBA Cross tabulation

		Experience of DBA			Total		
		1-5	6-10	11-15			
DBA In Organizations	1	Count	1	6	0	7	
		Expected Count	1.8	4.7	.6	7.0	
		% within DBA In Organizations	14.3%	85.7%	.0%	100.0%	
	2	Count	2	1	0	3	
		Expected Count	.8	2.0	.3	3.0	
		% within DBA In Organizations	66.7%	33.3%	.0%	100.0%	
	3	Count	0	1	1	2	
		Expected Count	.5	1.3	.2	2.0	
		% within DBA In Organizations	.0%	50.0%	50.0%	100.0%	
Total		Count	3	8	1	12	
		Expected Count	3.0	8.0	1.0	12.0	
		% within DBA In Organizations	25.0%	66.7%	8.3%	100.0%	

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DBA In Organizations * Experience of DBA	12	100.0%	0	.0%	12	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.869 ^a	4	.064
Likelihood Ratio	7.442	4	.114
Linear-by-Linear Association	.590	1	.442
N of Valid Cases	12		

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.869 ^a	4	.064
Likelihood Ratio	7.442	4	.114
Linear-by-Linear Association	.590	1	.442
N of Valid Cases	12		

a. 9 cells (100.0%) have expected count less than 5. The minimum expected count is .17.

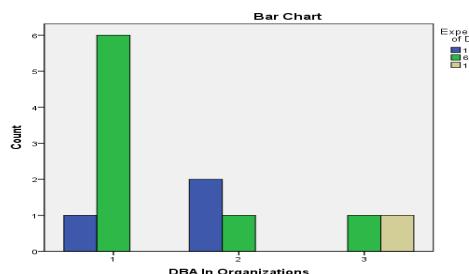


Figure18. DBA in Organizations

*Chi test associations between Data Recovery Method after data Lost * Level of Backup*

Research Hypothesis (H1) Data Recovery Method after data Lost and Level of Backup Cross Significance level=0.05%

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Data Recovery Method after data Lost * Level of Backup	12	100.0%	0	.0%	12	100.0%

Data Recovery Method after data Lost * Level of Backup Cross tabulation

		Level of Backup			Total
		full backup	offline backup	online backup	
Data Recovery Method after data Lost	backup	5	1	1	7
	Count				
	Expected Count	5.8	.6	.6	7.0
	% within Data Recovery Method after data Lost	71.4%	14.3%	14.3%	100.0%
	recovery method	2	0	0	2
	Count				
	Expected Count	1.7	.2	.2	2.0
	% within Data Recovery Method after data Lost	100.0%	.0%	.0%	100.0%
	both	3	0	0	3
	Count				
	Expected Count	2.5	.3	.3	3.0
	% within Data Recovery Method after data Lost	100.0%	.0%	.0%	100.0%
	Total	10	1	1	12
	Count				
	Expected Count	10.0	1.0	1.0	12.0
	% within Data Recovery Method after data Lost	83.3%	8.3%	8.3%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.714 ^a	4	.788
Likelihood Ratio	2.438	4	.656
Linear-by-Linear Association	1.041	1	.307
N of Valid Cases	12		

a. 8 cells (88.9%) have expected count less than 5. The minimum expected count is .17.

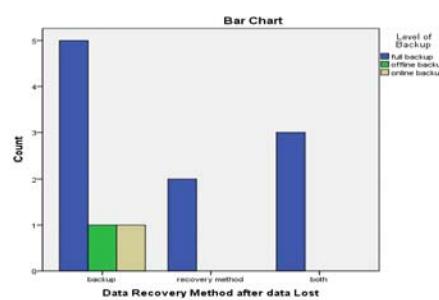


Figure19. Data recovery method after data lost

V. Conclusion

Big data related problems are faced by almost all organizations in the world. But I was survey only twelve organizations in Pakistan and create a result using SPSS software. In my report almost results are significant and Ho is rejected.

Many organizations still used olds methods. Some organizations have no knowledge about big data. In Pakistan 75% organizations saved their data using computer system. But yet have no idea about big data usage. In Pakistan 16.7% organizations using Hadoop, 8.3 using Jaspersoft and 8.3% using Talend Open Studio and 66.7% organizations still not using big data tools for saving data. They are all using others strategies for saving their data. 25% organizations using desktop as a generating source for data recording and 75% organizations are using both desktop and laptop for data recording. 16.7% organizations using archive file, 8.3% using default compression, 33.3% using data extractions tools and 41.7% using others strategies for information extracting and changing. Methods percentage for data cleaning which is used by organizations in Pakistan. 25% organizations using data mining tools, 16.7% using batch processing, 33.3% using others tools and 25% using no tools for data cleaning. 50% organizations using SQL query, 41.7% are using both SQL and PLSQL and 8.3% using no methods for querying data. 25% organizations using WEKA tools and 75% using others tools for mining their data.

ACKNOWLEDGEMENTS

The support provided by the Department of Information Technology, Govt. College University Faisalabad, Pakistan, is appreciatively acknowledged. Authors would also like to express thanks my respected teacher, for her precious suggestions and criticisms. In addition, authors would like to thanks my entire family members, who support me in my study and research.

REFERENCES

- [1] S. Kaisler, F. Armour, J. Alberto Espinosa, W. Money, (2013). 46th Hawaii International Conference on System Sciences. "Big Data: Issues and Challenges Moving Forward", Journal of IEEE, pp.645.
- [2] Hashem, Ibrahim Abaker Targio, et al (2015). "The rise of "big data" on cloud computing: Review and open research issues." *Information Systems* 47 pp.98-115.
- [3] Mervis, J. (2012). "Agencies Rally to Tackle Big Data", *Science*, pp.336 (4):22, June 6.
- [4] Oracle and FSN, (December 2012). "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity".
- [5] Rubinstein, Ira S., (2012). "Big Data: The End of Privacy or a New Beginning?" New York University Public Law and Legal Theory Working Papers.Paper pp.357.
- [6] Reips, Ulf-Dietrich; Matzat, Uwe Mining (2014). "Big Data" using Big Data Services". *International Journal of Internet Science*.1 (1):pp. 1-8.
- [7] Torgerson, P R; Torgerson, D (2010). Public health and bovine tuberculosis: what's all the fuss about? *Trends in Microbiology*, 18(2):pp.67-72.
- [8] Alexandru Adrian TOLE, (2013). Romanian American University, Bucharest, Romania, "Big Data Challenges "Database Systems Journal vol. IV, no. 3.
- [9] Katal. A, Wazid. M and R H Gouder, (2013). "Big Data: Issues, Challenges, Tools and Good Practices".pp.978- IEEE.
- [10] Suthaharan, S. (2014). "Big data classification: Problems and challenges in network intrusion prediction with machine learning". *Performance Evaluation Review*, 41(4), pp.70-73.
- [11] Tilmann. R, Sadoghi. M and Hans-Aron.J, (21 Aug 2012). "Solving Big Data Challenges for Enterprise Application Performance Management.ar".
- [12] Rubinstein, Ira S., (2012). "Big Data: The End of Privacy or a New Beginning?" .New York University Public Law and Legal Theory Working pp. 357.
- [13] Boyd, Dana; Crawford, Kate, (September 21, 2011). "Six Provocations for Big Data". *Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the InternetandSociety*.
- [14] Daniel, Ben, (2015). "Big data and analytics in higher education: Opportunities and challenges." *British journal of educational technology* 46.5 pp.904-920.
- [15] Demchenko.Y and Membrey. P, (2013) "Defining Architecture Components of the Big Data Ecosystem". System and Network Engineering Group University of Amsterdam, The Netherlands e-mail: {y.demchenko, C.T.A.M.deLaat}@uva.nl, Hong Kong Polytechnic University Hong Kong SAR, China e-mail: cspmembrey@comp.polyu.edu.hk.
- [16] Y. Chen, S. Alspaugh, R. Katz, (August 27th - 31st 2012). "Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of Map Reduce Workloads", Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, pp. 12.
- [17] S.S Sahoo, C. Jayapandian, G. Garg, F. Kaffashi, S. Chung, A. Bozorgi, (2013). "Heart beats in the cloud: distributed analysis of electrophysiological 'Big Data' using cloud computing for epilepsy clinical research." *Journal of the American Medical Informatics Association* 21.2: pp.263-271.
- [18] Raghupathi, Wullianallur, and Viju Raghupathi, (2014). "Big data analytics in healthcare: promise and potential." *Health information science and systems* pp.2.1, (2014): 3.
- [19] Liu, Yao, Peng Ning, and Michael K. Reiter, (2011). "False data injection attacks against state estimation in electric power grids." *ACM Transactions on Information and System Security (TISSEC)* pp.14.1, (2011): 13.
- [20] Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao, Hung-An, (2013). "Recent Advances and Trends in Predictive Manufacturing Systems in Big Data Environment". *Manufacturing Letters*. 1 (1): pp.38-41.
- [21] Magoulas, Roger; Lorica, Ben, (February 2009). "Introduction to Big Data". *Release 2.0*. Sebastopol CA: O'Reilly Media (11).
- [22] Andrea De Mauro, Marco Greco and Michele Grimaldi, (2015). "What is Big Data? A Consensual Definition and a Review of Key Research Topics", AIP Conference Proceedings 1644, 97.
- [23] Al-Rodhan, Nayef (2014-09-16). "The Social Contract 2.0: Big Data and the Need to Guarantee Privacy and Civil Liberties Harvard International Review". *Harvard International Review*. Retrieved pp.04-03.
- [24] Gandomi, Amir, and Murtaza Haider, (2015). "Beyond the hype: Big data concepts, methods, and analytics." *International Journal of Information Management*35.2:pp.137-144.
- [25] Haiping Lu, K. N. Plataniotis and A. N. Venetsanopoulos, (Jul. 2011). "A Survey of Multilinear Subspace Learning for Tensor Data", *Pattern Recognition*, vol. 44, no. 7, pp. 1540-1551.
- [26] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, (December 2012). "Business Intelligence and Analytics: From Big Data to Big Impact". *MIS Quarterly* Vol. 36 No. 4, pp. 1165-1188.

- [27] Nunan, Daniel and Di Domenico, (2015). "M. Market research the ethics of big data". International Journal of Market Research 55 (4), pp. 505-520. ISSN 1470-7853.

Enhancement of degraded Document Images using Retinex and Morphological Operations

Chandrakala H T
Research Scholar
Dept. of CSE
VTU Regional Research Center
Bengaluru, India
chandrakl80@gmail.com

Thippeswamy G
Professor and Head
Dept. of CSE
BMS Institute of Technology
Bengaluru, India
swamy.gangappa@gmail.com

Sahana D Gowda
Professor and Head
Dept. of CSE
BNM Institute of Technology
Bengaluru, India
sahanagowda@rediffmail.com

Abstract— Ancient historical inscriptions collected from various sources are image captured and stored as document images in digital libraries. Due to various factors, such as aging, degradation, erosion and deposition of foreign bodies on the inscriptions the quality of images captured is poor. These images are not ready for further processing such as reading, translation and indexing. Image Enhancement is an important phase for such document images before extracting information. A novel hybrid enhancement process has been proposed in this paper to highlight the text in the inscription, to make it more suitable for recognition using an OCR (Optical Character Recognition) system. The proposed method is a combination of Frankle McCann Retinex approach and Morphological operations which highlight the image contours by suppressing the background deformation and noise. The method is tested on a dataset of 300 camera captured estampage images of stone inscriptions written in ancient Kannada script. Experimental results show the efficacy of the proposed method.

Keywords- Frankle McCann Retinex; Thickening; Filling; Inscription images

Introduction

Inscriptions carved on stone, palm leaves, metal and shells are the historical documents which serve as the solitary and authentic records for understanding ancient history. These recorded experiences are useful in countless ways for the study and reconstruction of the social, economic, cultural, dynastic and political history of the mankind. Preservation of these documents is irrefutable if they must continue to serve as a reference in making further discoveries about the world. Unfortunately, these copies are at a serious risk of loss and extinction as they are deteriorating due to aging, natural disasters, risky handling, depositions and harsh weather

conditions. To preserve these valuable archaeological resources for future the Archaeological departments throughout the world excavate these inscriptions from their sources, create their Estampages and maintain a corpus of the same. But Estampages can also deteriorate in the long run due to breakage, aging, risky handling, dust and insects.

Digitization of these images is a more reliable solution for their preservation. Digitization creates faithful reproduction of Estampages in the form of digital images by either image capture or scanning. Digital images have longer shelf life and are easy to access and disseminate. Moreover, they can further take advantage of the power of digital image enhancement, possibilities of structured indexes, machine recognition and translation, mathematics of compression and communication. These technological solutions are very much needed to motivate the Archaeological Departments to convert their repository of historical documents into a digital library and to automate information extraction from these documents. Moreover, these digital documents are more readily accessible to historians and researchers compared to the originals that are not easily available for public viewing.

However, these digital images would be inherently degraded as they are captured from a source which is already deteriorated. Therefore, to make them suitable for automatic machine recognition and translation it is inevitable to pre-process them using suitable image enhancement technique. Image Enhancement improves the perception of information in an image for human viewing and for further automated image processing operations. The proposed work is the first effort to enhance the camera captured estampages of the stone inscriptions belonging to the 11th century Kalyani Chalukyan dynasty. These handwritten inscriptions are in Kannada script

and are collected from the corpus of the public organization-Archaeological Survey of India.

The techniques for image enhancement can be broadly classified as local and global methods. Global approach is an overall enhancement approach where the entire image is modified as per the statistics of the whole image. But meanwhile the smaller details are lost because the number of pixels in these small areas has no influence on the computation of global transformation. Whereas local enhancement can enhance even the smaller details in the image as it uses a small rectangular or square neighborhood with the centre moving from pixel to pixel over the entire image. The centre pixel of the window is modified with a value calculated based on the statistics of the other pixels of the window. Local enhancement is preferable for inscription images since the separation between the foreground text and the background is not prominent. This paper presents one such enhancement approach based on Frankle McCann Retinex algorithm coupled with morphological processing. The rest of the paper is organized as follows: section I discusses related work, section II gives a detailed explanation of the proposed enhancement scheme, section III discusses the experimental results and discussion and section IV concludes the paper.

I. RELATED WORK

As available in the literature, the enhancement of Historical handwritten documents has been performed using Background light intensity normalization [29], directional wavelet transform [28], Background light intensity normalization [40] and Hyperspectral imaging [39]. They mainly address the issues like background noise and ink bleed through. Specific to inscription image enhancement median filtering technique [32-35] has been used extensively. Curvelet transform [30] and shearlet transform [31] in combination with morphological operations have been used to denoise south Indian palmscripts. Natural Gradient based Fast Independent Component Analysis technique has been employed to enhance stone inscriptions of Hampi [27]. But these techniques are not suitable for inscription estampage images as they might result in uneven contrast stretching.

Inscription images do not have clear visible difference between the foreground text and the background. Many times the deformation in the background would look like part of foreground text rendering poor visual appearance to these images. Retinex filtering [1, 3, 10, 11] is an enhancement method which compensates for non-uniform contrast by separating the illumination from the reflectance in a given image. It decreases the influence of the reflectance component, thus enhancing the original image to its true likeness. Hence it is more suitable for enhancement of Inscription document images. Although Retinex methodology had been used so far to enhance medical images[22], satellite images[21], natural scene images[5], nighttime images[20] and many more, it was only used for skew correction of document images[23] till [41] used it for contrast enhancement of inscription document images. The proposed enhancement scheme aims at improving the contrast enhancement results achieved by [41].

The Retinex algorithms published in the literature can be classified into four categories: Path based algorithms, Recursive algorithms, Center Surround algorithms and Variational algorithms. In path based algorithms the value of the new pixel depends on the product of ratios along the stochastic paths [11]-[14]. Recursive algorithms replace the path computation by a recursive matrix comparison [7]-[9]. These algorithms are computationally more efficient than the path based algorithms. In Center Surround method [3]-[5] a given pixel value is compared with the surrounding average pixel values to compute the new pixel. The variational Retinex algorithms [15]-[17] convert the constraints of illumination and reflectance into a mathematical problem and then obtain the new pixel value by solving equations or optimization problems. Morphological operations have been traditionally used as an effective tool for noise removal and enhancement of digital images [24-26].

Frankle McCann Retinex [7] algorithm, a recursive variant of traditional Retinex was found to be more suitable to highlight the text contours in the inscription image as it can stretch the image contrast simultaneously compressing the dynamic range rendering better visual clarity. Following which some morphological operations can be applied to suppress background noise and deformation. The following section provides the details of these techniques employed in the proposed method.

II. METHODOLOGY

The proposed enhancement scheme integrates Frankle McCann Retinex algorithm with Morphological Processing to enhance the inscription document images. Frankle McCann Retinex (FMR) algorithm [7] performs pixel level contrast stretching rendering sharp contrast to the entire image. To highlight the text contours and to suppress background noise and deformation, Morphological operations are performed on the Retinex enhanced images. The following subsections explain the FMR enhancement and the Background noise suppression in detail.

A. Frankle McCann Retinex

The principle of Frankle McCann algorithm [7] is shown in figure 1 below:

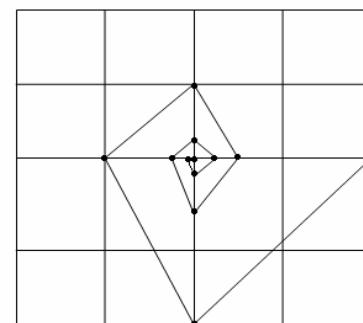


Figure 1: Principle of Frankle McCann Retinex Algorithm

At each step, the comparison is implemented using the Ratio-Product-Reset-Average operation. The process continues

until the spacing decreases to one pixel. The Ratio-Product-Reset-Average operation is given by the equation:

$$r_p^{k+1} = (\text{reset}(i_p - i_q + r_q^k) + r_p^k) / 2 \quad (1)$$

Where p is the neighborhood center pixel index and q is the index of one of the neighboring point. Let $p=(x,y)$, then $q \in \{(x \pm d^k, y), (x, y \pm d^k)\}$, where d^k is the shift distance corresponding to the k-th update operation. In the iterative procedure, for a given p, q is spirally taken, and d^k is progressively reduced towards zero.

Since the operations are performed in logarithmic domain, the term is the ratio between the original intensity at p and that at q. The following addition is the product operation. Then the ratio-product term is reset to a constant whenever it exceeds the constant. And finally the reflectance estimation is updated by averaging the last estimation and the reset term.

B. Morphological Operations

The background noise pixels in the image produce artificial edges which are also enhanced by Retinex processing. These unwanted background edges interfere with the foreground text which hampers their visual clarity. In order to suppress these artifacts some morphological operations are performed on the binarized FMR output. The foreground text is accumulated into a connected component by applying Morphological Thickening, Filling and Bridging.

Thickening is the morphological dual of thinning. It is defined as

$$A \circ B = A \cup (A \otimes B) C \quad (2)$$

where A is the image matrix (set),

B is a structuring element suitable for thickening

\otimes is the hit or miss transformation operation

T

The morphological fill operation fills all the holes with ones for binary images. The hole filling algorithm first generates an array X_0 . X_0 contains all zeros except at the location corresponding to the given point in each hole, which is set to one. This is followed by the following procedure:

$$X_k = (X_{k-1} \oplus B) \cap A^C \quad k=1,2,3,\dots \quad (3)$$

where B is the symmetric structuring element

\oplus is the dilation operation

The algorithm terminates at the iteration step k if $X_k = X_{k-1}$. The set X_k then contains all the filled holes; the union of X_k and A contains all the filled holes and their boundaries. The dilation would fill the entire area if left unchecked. However, the intersection at each step with the complement of A limits the result to inside the region of interest.

Bridging operation ties the unconnected pixels in the binary image by setting all zero valued pixels to one if they have two nonzero neighbours that are not connected. The objects signified by the connected components are assigned labels. For each object pixel summation is performed. If this sum is higher than the assumed threshold then the object is detected as valid

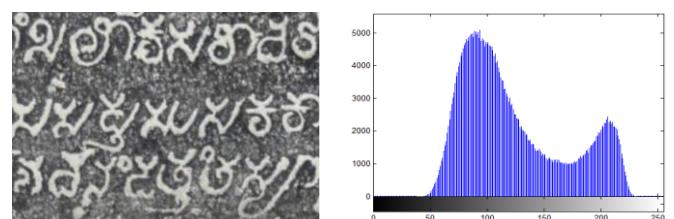
text area and put into a mask image M. Finally the Region of Interest(ROI), that is the foreground text is extracted by computing Hadamard Product of X and M given by:

$$(X \circ M)_{i,j} \leftarrow (X)_{i,j} (M)_{i,j} \quad (4)$$

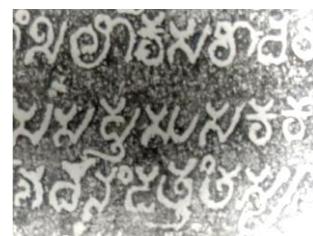
III. RESULTS AND DISCUSSION

The proposed Enhancement scheme is tested on the dataset of 300 camera captured images of ancient Kannada inscription Estampages that belong to the Kalyani Chalukyan era of 11th century. The images are captured using a camera of 13 Megapixel resolution. The visual quality of the original image as shown in Fig 2(a) is poor as it is infected by background noise and interference of the background pixels with the foreground text. Three different types of Retinex techniques namely Single Scale Retinex(SSR), MultiScale Retinex(MSR) and Frankle McCann Retinex (FMR) and the proposed method (FMR with Morphological operations) were tried on the dataset. SSR method enhanced the text but output resulted in significant greying out effect in some parts of the image as shown in Fig 2(b) leading to loss of some visual content. Quite a similar effect was observed with MSR as shown in Fig 2(c).

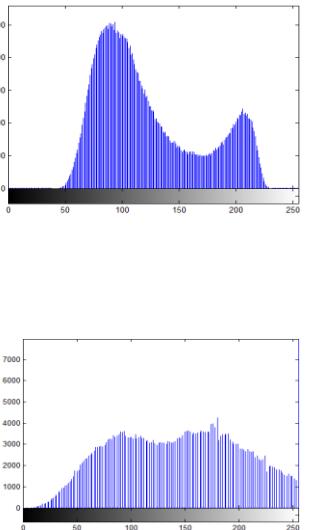
Frankle McCann Retinex algorithm is applied on the logarithmic version of the original image. The corresponding output is shown in Fig 2(d). Though the foreground text looks enhanced when compared to Fig 2(a), the overall image suffers from slight greying; a consequence of Retinex algorithm, also the unwanted artifacts gets enhanced. This might result in less accuracy of Optical Character Recognition to be performed later. So, to further enhance this result Morphological Processing is performed on the FMR output and the result is shown in Fig 2(e). Morphological Processing has removed the greyish look of the image and has improved the contrast by removing the unwanted background pixels making it more suitable for further processing steps.



(a)



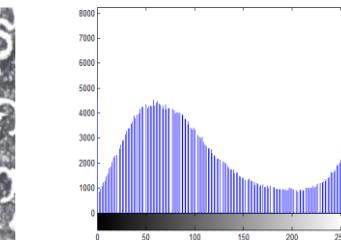
(b)



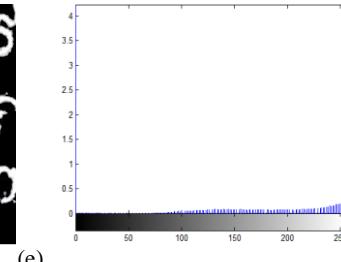
(c)



(c)



(d)



(e)

Figure 2 : The enhancement results achieved using the different Retinex approaches with the corresponding histogram plots (X-axis represents pixel intensity and Y-axis represents pixel count). (a) Original image (b) SSR (c) MSR (d) FMR (e) FMR with Morphological processing

Experimentation was done on our estampage dataset of 11th century Kannada stone inscriptions and also on the standard Handwritten text datasets-HDIBCO 2010, HDIBCO 2014 and HDIBCO 2016 to give a comparative study. The quality of the enhancement results were evaluated by measuring their Standard Deviation and Root Mean Square (RMS) Contrast.

A. Standard Deviation

$$STD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2} \quad (5)$$

where X_i is a one dimensional array of N pixel intensities of the given image and \bar{X} is the corresponding mean given by:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (6)$$

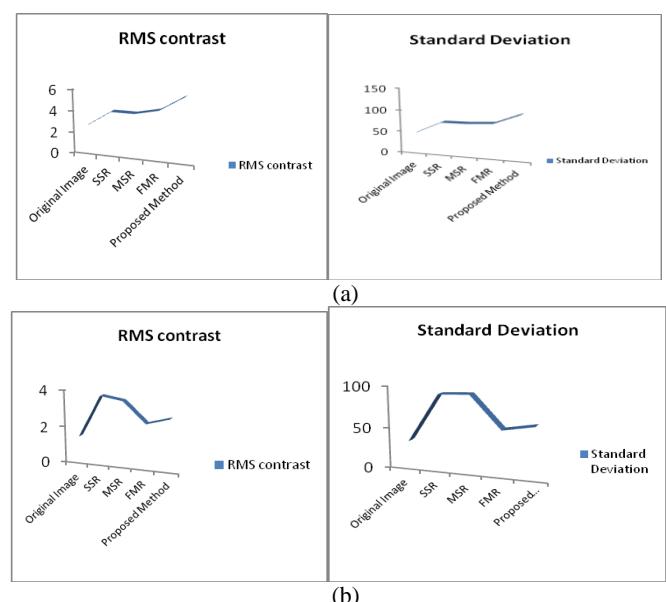
B. RMS Contrast

$$RMS\ contrast = \sqrt{\frac{1}{MN} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} (I_{ij} - \bar{I})^2} \quad (7)$$

where I_{ij} is an image of size $M \times N$ whose pixel intensities are normalized in the range [0,1]. \bar{I} is the mean intensity of all pixel values in the image I_{ij} .

TABLE 1: RMS contrast and Standard Deviation values achieved using SSR, MSR, FMR and the proposed method on Estampage dataset. These methods are evaluated on the standard Handwritten DIBCO datasets- HDIBCO2010, HDIBCO2014 and HDIBCO2016 .

Dataset	Measures	Original Image	SSR	MSR	FMR	FMR with Morphological processing
Estampage	RMS contrast	2.61	3.99	3.98	4.44	5.78
	Standard Deviation	45.82	72.80	73.77	77.89	101.41
HDIBCO2010	RMS contrast	1.18	3.56	3.66	2.14	2.44
	Standard Deviation	23.36	80.27	83.58	43.08	46.76
HDIBCO2014	RMS contrast	1.41	3.74	3.54	2.40	2.75
	Standard Deviation	31.79	91.47	93.42	54.11	61.12
HDIBCO2016	RMS contrast	2.43	6.09	6.21	3.31	3.52
	Standard Deviation	38.68	72.30	75.86	52.46	56.34



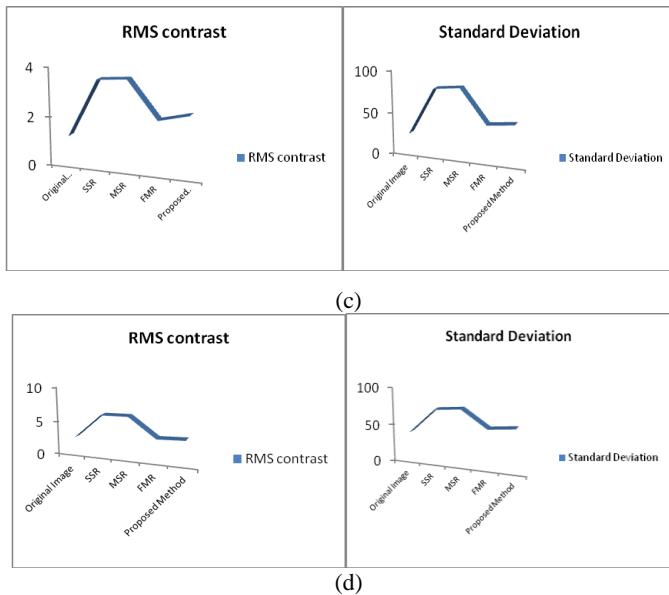


Figure 3 : RMS contrast and Standard Deviation plots on different Datasets (X-axis represents pixel intensity and Y-axis represents enhancement method).
(a) Estampage (b) HDIBCO 2010 (c) HDIBCO 2014 (d) HDIBCO 2016

IV. CONCLUSION AND FUTURE SCOPE

Based on the degradation characteristics of Inscription estampage images an improved enhancement approach which integrates Morphological Processing with Frankle McCann Retinex algorithm has been implemented. This scheme highlights the text by iterative contrast stretching and suppresses the background artifacts through mathematical morphology . The results thus achieved show superior visual clarity with the best Standard Deviation and RMS contrast when compared to the traditional Retinex variants. However it was observed that the proposed method took too much of computational time. This is one issue that can be addressed in future.

REFERENCES

- [1] Edwin H Land , “The Retinex Theory of Color Vision”, J. Scientific American, Vol 237 No 6 P108-128. 1997
- [2] Ana Belen Petro, Catalina Sbert, Jean-Michel Morel ,” Multiscale Retinex ”, Image Processing Online (IPOL), ISSN 2105-1232. 2014
- [3] Z. Rahman, D. J. Jobson, G. A. Woodell, “ Retinex processing for automatic image enhancement”, Human Vision and Electronic Imaging VII, SPIE Symposium on Electronic Imaging, Proc. SPIE 4662, 2002
- [4] D J Jobson, Z Rahman, G A Woodell, “Properties and performance of a center/surround Retinex”, IEEE Trans.Image Processing, Vol 6, no.3, p. 451-462.1997
- [5] D J Jobson, Z Rahman, G A Woodell, “A multiscale Retinex for bridging the gap between color images and the human observation of scene”s, IEEE Trans. Image Processing, Vol 6, no.7, p. 965-976.1997
- [6] R. C. Gonzalez, R. E. Woods, “Digital Image Processing”, 2nd ed., New Jersey: Prentice-Hall. 2002
- [7] Frankle J , McCann J, “Method and apparatus for lightness imaging”, US,4384336[P]. 05-17. 1983
- [8] Funt B,Ciurea F,McCann J, “Retinex in Matlab”, Journal of Electronic Imaging ,13(1):48-57. 2004
- [9] J McCann, “Lesson learned from mondrians applied to real images and color gamuts”, in Proc. IST/SID 7th Color Imag.Conf. ,pp. 1-8.1999
- [10] Lei Ling ,Zhou Yingqin,Li Jingwen, “An investigation of Retinex Algorithms for Image Enhancement”, Journal of Electron(China) ,Vol.24 No.5. 2007
- [11] E Land J McCann, “Lightness and Retinex theory”, J. Optical Society of America ,vol.61,no.1 pp 1-11.1971
- [12] E H Land, “Recent advances in Retinex theory”, Vis.Res., vol.26, no.1,pp 7-21.1986
- [13] E .Provenzi, L D Carli, A Rizzi, D Marini, “Mathematical definition and analysis of the Retinex algorithm”, J Opt. Soc. Amer.vol 22,pp.2613-2621.2005
- [14] D Marini, “A computational approach to color adaptation effects”, Image Vis. Comput., Vol. 18,no.13, pp. 1005-1014.2000
- [15] A Blake, “Boundary conditions for lightness computation in Mondrian world”, Comput. Vis. Graph. Image Process., Vol. 32, pp. 314-327.1985
- [16] B Funt, M Drew , M Brockington, “Recovering shading from color Images”, in Proc. 2nd Eur. Conf. Comput. Vis., pp.124-132.1992
- [17] D Terzopoulos, ”Image analysis using multigrid relaxation methods”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. PAMI-8, No.2, pp. 129-139.1986
- [18] Shengdong Pan, Xiangjiing An, Hongtao Xue, Hangen He, “Improving Iterative Retinex Algorithm for Dynamic Range Compression”, Proceedings of 2nd International Conference and Information Application.2012
- [19] Jia Li, “Application of image enhancement method for digital images based on Retinex theory”, Journal Optik124,5986-5988, Elsevier.2013
- [20] Haoning Lin, Zhenwei Shi, “Multi scale Retinex improvement for nighttime image enhancement”, Journal Optik 125,7143-7148, Elsevier.2014
- [21] Akram hashemi Sejzei, Mansur Jamzad, “Evaluation of various digital image processing techniques for detecting critical crescent moon and introducing CMD- A tool for critical crescent moon detection”, Journal Optik 127,1511-1525, Elsevier.2016
- [22] Yifan Wang, Hongyu Wang, Chuanli Yin, Ming Dai, “Biologically inspired image enhancement based on Retinex”, Journal of Neurocomputing 177 373-384, Elsevier.2016
- [23] Marian Wagdy, Ibrahima Faye, Dayang Rohaya, “Degradation Enhancement for the Captured Document Image using Retinex Theory”, International Conference on Information Technology and Multimedia(ICIMU).2014
- [24] Hamid Hassanpour, Najmeh Samadiani, Mahdi Salehi, “Using morphological transforms to enhace the contrast of medical images”, The Egyptian Journal of Radiology and Nuclear Medicine, Elsevier. 2015
- [25] Cao Yuan, Yaqin Li, “Switching median and morphological filter for impulse noise removal from digital images”, Journal Optik126, 1598-1601, Elsevier.2015
- [26] Shijian Lu, Ben M Chen, C C Ko, “Perspective rectification of document images using fuzzy set and morphological operations”, Journal of Image and Vision Computing 23 541-553, Elsevier.2005
- [27] Indu Sreedevi, Rishi Pandey, N Jayanthi, Geetanjali Bhola , Santanu Chaudhary,”Enhancement of Inscription Images”, 978-4673-5952-8/13 , IEEE .2013
- [28] Qian Wang, Tao Xia, Lida Li, Chew Lim Tan, “Document Image Enhancement using Directional Wavelet”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1063-6919/03, IEEE .2003
- [29] Zhixin Shi , Venu Govindaraju, “Historical Document Image Enhancement using Background Light Intensity Normalization”, Proceedings of 17th International Conference on Patten Recognition, 1051-4651/04, IEEE .2004
- [30] B Gangamma, Srikantha Murthy K ,”A combined approach for degraded Historical Documents denoising using Curvelet and Mathematical Morphology”,978-1-4244-5967-4/10, IEEE.2010
- [31] Ranganatha D, Ganga Holi, “Historical Document Enhancement using Shearlet Transform and mathematical morphological operations”, 978-1-4799-8792-4/15, IEEE .2015

- [32] G Janani, V Vishalini, P Mohan Kumar, "Recognition and Analysis of Tamil inscriptions and mapping using Image Processing Technique's, 978-1-5090-1706-5/16 , IEEE .2016
- [33] Saleem Pasha, M C Padma, "Handwritten Kannada Character Recognition using Wavelet Transform and structural features", International Conference on Emerging Research in Electronics, CST, 978-4673-9563-2/15, IEEE .2015
- [34] G Bhuvaneswari, V Subbiah Bharathi, "An efficient algorithm for recognition of ancient stone inscription characters", 7th International Conference on Advanced Computing, 978-5090-1933-5/15, IEEE.2015
- [35] N Jayanthi, S Indu, P Gola, P Thripathi, "Novel method for manuscript and inscription text extraction", 3rd International Conference on Signal Processing and Integrated Networks, 978-4673-9197-9/16, IEEE. 2016
- [36] Shafali Gupta, Yadwinder Kaur. Review of different local and Global contrast enhancement techniques for digital image. International Journal of Computer Applications,0975-8875, Volume 100-No.18.2014
- [37] Wenye Ma, Jean-Michel Morel, Stanley Osher , Aichi Chien, "An L1-based variational model for Retinex theory and its application to medical images", Proceedings,IEEE Computer Society Conference on Computer Vision and Pattern Recognition 20-25.2011
- [38] Anu Namdeo , Sandeep Singh Bhadriya, "A Review on Image Enhancement Techniques with its Advantages and Disadvantages", International Journal for Science and Advance Research In Technology ISSN : 2395-1052 - Volume 2 Issue 5 .2016
- [39] Seon Joo Kim, Fanbo Deng, Michael S Brown, "Visual Enhancement of old documents with hyperspectral imaging",Journal of Pattern Recognition 44-1461-1469, Elsevier. 2011
- [40] Zhixin Shi, Venu Govindaraju, "Historical Document Immage Enhancement using Background Light Intensity Normalization", Proceedings of 17th International Conference on Pattern Recognition 1051-4651.IEEE.2004
- [41] Chandrakala HT, Thippeswamy G, "Epigraphic Document image enhancement using Retinex method", Proceedings of 3rd international symposium of signal processing and intelligent recognition systems, Book chapter in Advanced in signal processing and intelligent recognition systems, Springer,ISBN:978-3319679334, 2017.

AUTHORS PROFILE



Mrs.Candrakala H T is a university second rank holder in her post graduation in CSE from Visvesvaraya Technological University in 2012. Currently she is a pursuing her PhD in the field of Digital Image Processing under Visvesvaraya Technological University. She has 8 years of Teaching experience and 3 years of Research experience. She is now working as Assistant Professor in the Department of Computer Science,GFGCM,Tumkur University,India. Her research areas of interest include image processing, pattern recognition, computer vision and data mining. She is a life member of IEI and ISTE. She has published 10 research papers in reputed journals and conferences including springer .



Dr. Thippeswamy G received his ME in CSE from Bangalore University in 1997 and PhD in Digital Image Processing from Mangalore University in 2012. He has 24 years of Teaching experience and 8 years of Research experience. His research areas include image processing,pattern recognition, computer vision and data mining. He is now working as Professor and HOD in the Department of CSE ,BMS Institute of Technology, Bengaluru, India. He is a life member of CSI,IEI and ISTE. He has published more than 20 research papers in reputed journals and conferences including springer.



Dr. Sahana D Gowda received her ME in CSE from Bangalore University and PhD in Digital Image Processing from University of Mysore . She has 16 years of Teaching experience and 10 years of Research experience. Her research specializations include image processing,pattern recognition, computer vision and data mining. She is now working as Professor and HOD in the Department of CSE , BNM Institute of Technology, Bengaluru, India. She is a life member of CSI and ISTE. She has published more than 20 research papers in reputed journals and conferences including IEEE and springer .

COMPARATIVE ANALYSIS OF K-MEANS DATA MINING AND OUTLIER DETECTION APPROACH FOR NETWORK-BASED INTRUSION DETECTION

¹ Lazarus Kwao, MPhil IT, Dept. Of Computer Science, KNUST, Kumasi, Ghana,
lazoe16@yahoo.com

² Joseph Kobina Panford, Dept. Of Computer Science, KNUST, Kumasi, Ghana,
jpanford@yahoo.com

³ James Ben Hayfron-Acquah, Dept. Of Computer Science, KNUST, Kumasi, Ghana,
jbha@yahoo.com

Abstract - New kind of intrusions causes deviation in the normal behaviour of traffic flow in computer networks every day. This study focused on enhancing the learning capabilities of IDS to detect the anomalies present in a network traffic flow by comparing the k-means approach of data mining for intrusion detection and the outlier detection approach. The k-means approach uses clustering mechanisms to group the traffic flow data into normal and abnormal clusters. Outlier detection calculates an outlier score (neighbourhood outlier factor (NOF)) for each flow record, whose value decides whether a traffic flow is normal or abnormal. These two methods were then compared in terms of various performance metrics and the amount of computer resources consumed by them. Overall, k-means was more accurate and precise and has better classification rate than outlier detection in intrusion detection using traffic flows. This will help systems administrators in their choice of IDS.

Key Words: K-Means, Outlier Detection Approach, Intrusion Detection, Network- based, NOF, clusters

1. INTRODUCTION

1.1 Background of Study

An intrusion is a malicious or unauthorized attempt or activity to access, modify, control, or create an unreliable or unusable system (Anderson, 1980, and Jirapummin, 2000). Intrusions attempt, or intrusions can result from external or internal intruders. Today, it is difficult to maintain a high level of security to ensure secure and reliable communication of information between different organizations as the speed and complexity of networks increases rapidly, especially when these networks are open to the public. The number and types of intrusions have increased considerably. Secure communication via computer networks and other systems carries the risk of intrusion and abuse. Thus, intrusion detection systems (IDS) have become a necessary part of the security of computers and networks (Hoque, Mukit, Bikas and Naser, 2012). Intrusion Detection Systems addresses three critical security functions: detection, observation and response to illegal actions by intruders. The intrusion detection system (IDS) is used as the second defense in the computer and in the network system to ensure security (Bijone, 2016). An intrusion detection system does not prevent an interruption, detects, observes and informs a system administrator. This response typically incorporates attempts to contain or maintain such damage, for example, when closing a network connection. When an IDS detects illegal system activity, it logs these events, stores important data, activities, alerts, and the administrator through a warning and, in some cases, attempts to intervene. In addition to the undeniable benefits of an IDS, the archived data and recordings provide satisfactory scientific evidence and can be used as evidence in a legitimate legal case against the intruder.

1.2 Statement of the Problem

There are intruders trying to get unauthorized access to network systems day after day (Sundaram, 1996). Meanwhile, there are problems such as identifying new attacks when it comes to intrusions entering the network when a large amount of data is available. Therefore, adequate training is needed for the IDS to know new types of attacks very frequently. This study proposes a new technique that detects and reduces the amount of time and resources required by the learning algorithm in a network traffic flow, comparing the effectiveness of the k-means approach data mining for intrusion detection and outlier detection using the nearest neighborhood factor.

1.3 Research Objectives

The performance of these two approaches is compared across multiple metrics of confusion and performance metrics and an analysis is performed to determine which of the two approaches is most effective for intrusion detection in network traffic flows.

The anomalies present in the traffic flow are of following types:

- Protocol anomaly (e.g., HTTP traffic on a non-standard port).
- Statistical anomaly (e.g., too much UDP compared to TCP traffic).
- Hybrid anomaly (combination of the above).

1.4 Research Questions

The following are the research questions that were posed to accomplish the objectives.

- Using the performance metrics which of the two approaches is effective for Intrusion detection?
- What is the amount of computer resources consumed by the two approaches?

1.5 Significance of the Study

This study will be significant in providing an in-depth understanding of the performance of the k-means data mining algorithm and outlier detection for intrusion detection. The study will also compare the amount of computer resources (CPU and RAM) consumed by outlier detection and k-means and come out with the one which is time-expensive.

2. LITERATURE REVIEW

The extremely connected computing world has equipped intruders and hackers with new techniques for unauthorized activities. The cost of such temporary or permanent damages caused by their activities to computer systems have entreated organizations to increasingly implement several structures to monitor information flow in their networks (Chandel, 2017). Several security methods have been developed to counter these security threats at different levels of the Transport Control Protocol/ Internet Protocol (TCP/IP) protocol stack. These security threats in the form of intrusions are generally hidden in nature and enter the network through packets or flows. To counter such threats, an intrusion detection system is required to alert the network administrators of a possible attack. There are two fundamental strategies to the planning of IDSs. In an exceedingly misuse detection-based IDS, intrusions are detected by examining and exploring events that correspond to established signatures of intrusions or vulnerabilities.

Besides, an anomaly detection based mostly IDS detects intrusions by observing for unusual network traffic.

Related Works

K-Means in Intrusion Detection

This major work is done in the areas such as usage of k-means to partition the data, categorizing botnets using k-means, usage of k-means in detecting intrusions in networks.

Bohara, Thakore and Sanders (2016) agreed on a method to carry out intrusion detection using k-means to partition the data as unsupervised learning approach. They have proposed new distance metrics which can be used in the k-means algorithm to carefully relate the clusters. They have partitioned data into more than one cluster and correlated them with known behaviour for evaluation. Their results have proven that k-means clustering is a better method to categorizing the data using unsupervised techniques for intrusion detection when several types of datasets are available.

As clustering algorithm proves to be very beneficial having large unlabelled dataset, Raykov, et al. (2016) provided an entire analysis of the NSL-KDD dataset and the attacks provided. They used k-means algorithm for this purpose and additionally represented the distribution of instances in clusters imparting better illustrations of the instances and making it clearer to apprehend.

In (Lisehroodi, Muda & Yassin. 2013), they proposed a hybrid framework based totally on neural community Multilayer perceptron (MLP) and K-means Clustering.

Wang Shunye et. al (2013) proposed enhanced k-means clustering algorithm basically consists of three steps. The first step talks about the construction of the dissimilarity matrix. Secondly, Huffman algorithm is used to create a Huffman tree according to dissimilarity matrix. The output of Huffman tree gives the initial centroids. Finally, the k-means clustering algorithm is appropriate to initial centroids to get k cluster as output. Wine and Iris datasets are selected from UIC machine learning repository to test the enhanced algorithm. Proposed algorithm gives better accuracy rates and results than the traditional k-means clustering algorithm.

Md.SohrabMahmud et. al (2012) proposed an algorithm uses heuristic method to calculate initial k centroids. The proposed algorithm yields accurate clusters in lesser computational time. The proposed algorithm initially calculates the average score of each data objects that has multiple attributes and weight factor. Next, the Merge sort is applied to arrange the output that was generated in first phase. The data points are then divided into k cluster. Finally, the nearest possible data point of the mean is taken as initial centroid. Although the proposed algorithm still deals with the problem of assigning number of desired k-cluster as input.

Juntao Wang and Xiaolog (2011) in his study, an improved k-means clustering algorithm to deal with the problem of outlier detection of traditional k-means clustering algorithm. The enhanced algorithm makes use of noise data filter to deal with this problem. Outliers can be detected and removed by using Density based outlier detection method. The purpose of this method is that the outliers may not be engaged in computation of initial cluster centres. The Factors used to test are clustering time and clustering accuracy. The drawback of the enhanced k-means clustering algorithm is that while dealing with large scale data sets, it takes more time to produce the results.

Munz, Li and Carle (2007) proposed a new method for data mining the use of k-means to sense intrusions. They have categorised the flow of data into clusters of normal and abnormal behaviour and have offered well-described facts of the intrusions and data flow.

However, researchers have additionally proposed few adjustments in the k-means algorithm to make it adaptable to the type of datasets with unique kind of networks. This proposal of k-means algorithm offered, makes it appropriate to be referenced for studies related to k-means.

Outliers in Intrusion Detection

Chandola., et, al (2009) defines outlier mining as the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.

Jabez& Muthukumar (2015) explained a new concept for intrusion detection in computer networks using distance and density-based approaches. They have introduced a metrics called as neighborhood outlier factor which is used to measure the anomaly dataset.

Manandhar and Aung (2014) have proposed an anomaly-based IDS using outlier detection. They have used the normal data instances to build a base model and declare all other data instances which do not obeys the base model.

Gosavi and Wadne (2014) compares various unsupervised outlier detection approaches using various techniques of outlier detection such as local outlier factor (LOF) and local distance-based outlier factor (LDOF).

Kriegel, Kroger and Zimek (2010) has focused on few other approaches to outlier detection which includes model-based approach, proximity-based approach and angle based approach for intrusion detection.

Wu et al, (2007) proposed a new outlier mining algorithm based on index tree, named TreeOut, designed to detect the outliers. Outliers have the weight greater than the threshold. in this technique the upper and lower bound of the weight of each record is calculated for r-region and index tree to avoid needless distance calculation. This algorithm is straightforward to implement, and more appropriate to detect intrusions in the audit data. The outlier detection method is effective in reducing false positive rate with desirable and correct detection rate (Bhuyan, et. al., 2011). This work for the generation of outliers for detecting intrusions is as follows.

3. METHODOLOGY

The methodology for this research was Design Science Research using simulation of scenarios. To help with this simulation, Graphical NetworkTrafficView and Wireshark was employed and the architecture for the simulation scenarios is illustrated in Figure1. Graphical NetworkTrafficView and Wireshark were chosen because they have user-friendly Graphical User Interface (GUI) and enables users to capture live network traffic packets in real time and displays general statistics about your network traffic. MATLAB was used for the analysis of the data captured by NetworkTrafficView and Wireshark. Matlab was used to do the performance analysis of the k-means and outlier detection approaches.

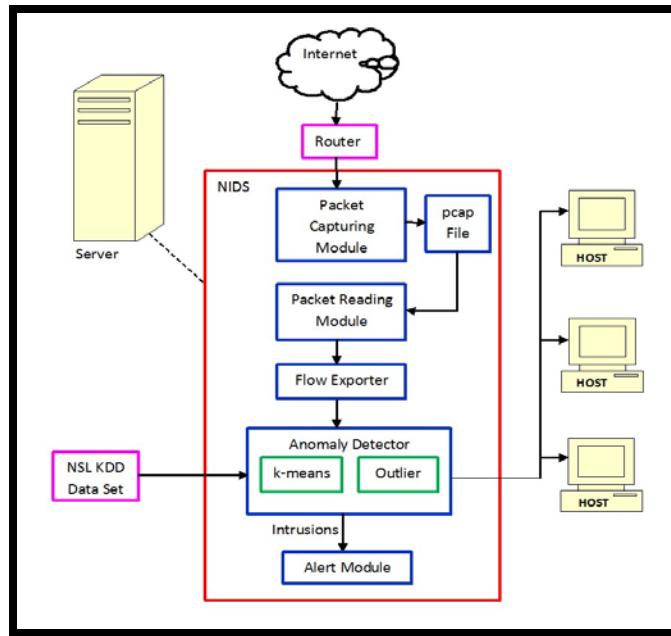


Figure 1: System architecture of IDS (Chandel, 2017)

Input data

The process starts with the collection of input data and clustering them into flow records by the IDS. This input data containing real internet traffic is collected at four times with different types of traffic characteristics. This input data was captured from the network setup of few tertiary institutions connected to each other to share knowledge and is used by students of the tertiary institutions to access the internet. Each set of data captured forms a dataset consisting of many flow records. The study has captured a smaller set of traffic ranging between 1-3 minutes for each dataset for analysis. Each dataset consists of varying both normal and attack data. This attack data consists of mainly TCP injection, UDP flooding and ICMP flooding attacks. The most important fields of a flow record are the total number of packets in a flow and the total number of bytes in each flow. This combination of attributes of the flow helps in detecting anomalies in the total amount of traffic. Another combination of attributes in the flow to detect the anomalies are the sources and destination IPs and port pairs which provide input to detect the port scans. Table 1 shows the attributes of packets and flows contained in the datasets.

Table 1: Input Data Attributes

Dataset	TCP Packets	UDP Packets	ICMP Packets	Total Flows
Dataset 1	3749	1185	220.5	5154
Dataset 2	1616	1610	345	3570
Dataset 3	2388	657	64.5	3110
Dataset 4	11034	1224	231	12489

4. ANALYSIS

The performance metrics are evaluated for the two approaches and a comparative study is presented

Table 2: Baseline characteristics results.

Dataset	Normal Traffic	Abnormal Traffic
Dataset 1	55.31%	44.69%
Dataset 2	58.50%	41.50%
Dataset 3	53.39%	46.61%
Dataset 4	55.60%	44.40%

Results of Experiment- K-Means Evaluation

The data captured in the four datasets as mentioned in table 2 was the input data for the k-means algorithm. k-means algorithm clustered the traffic into normal or abnormal flows. Table 3 shows the percentage of the traffic clustered into normal or abnormal flows for each dataset using k-means. The study plotted the normal and abnormal clusters for each dataset using k-means as shown in figure 1-4. Each flow record of both the clusters are marked as green for normal flow and red for abnormal flow as shown in figures 1-4.

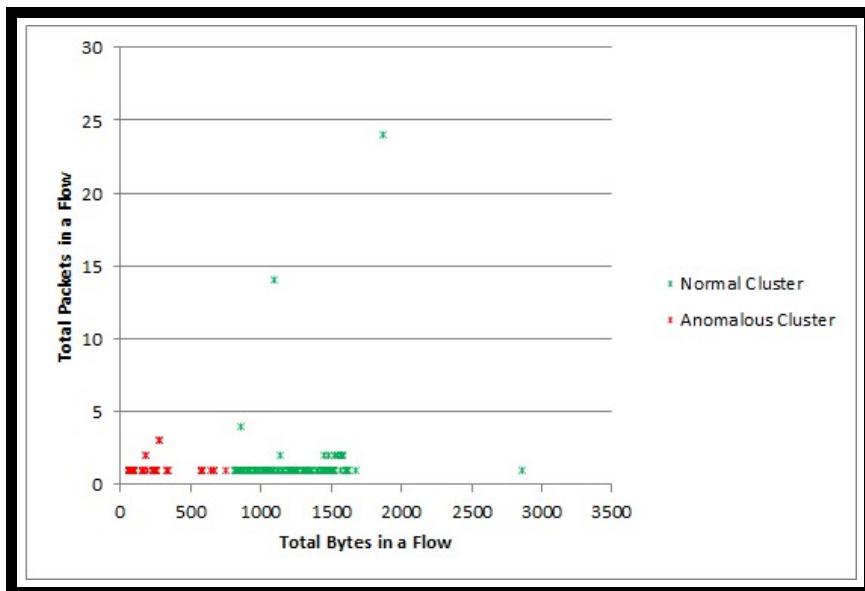


Figure 1: k-means clustering on dataset 1.

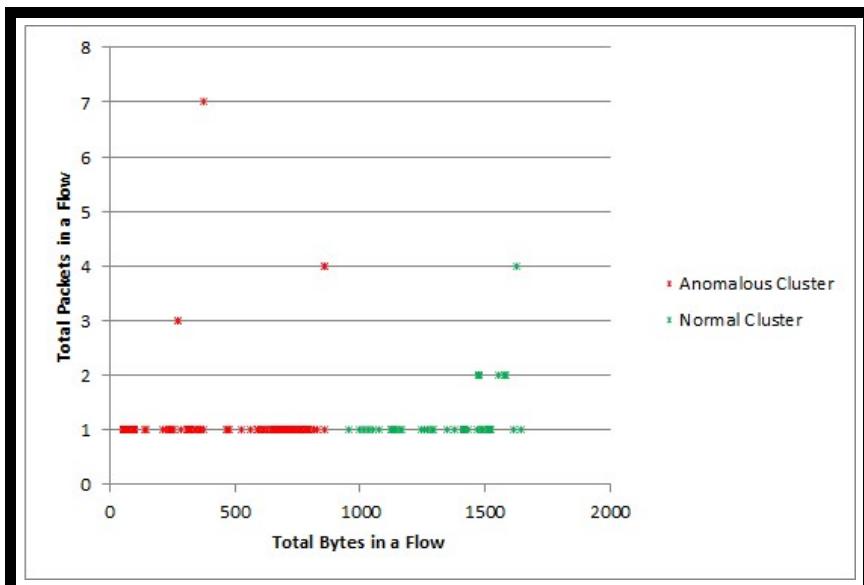


Figure 2: k-means clustering on dataset 2.

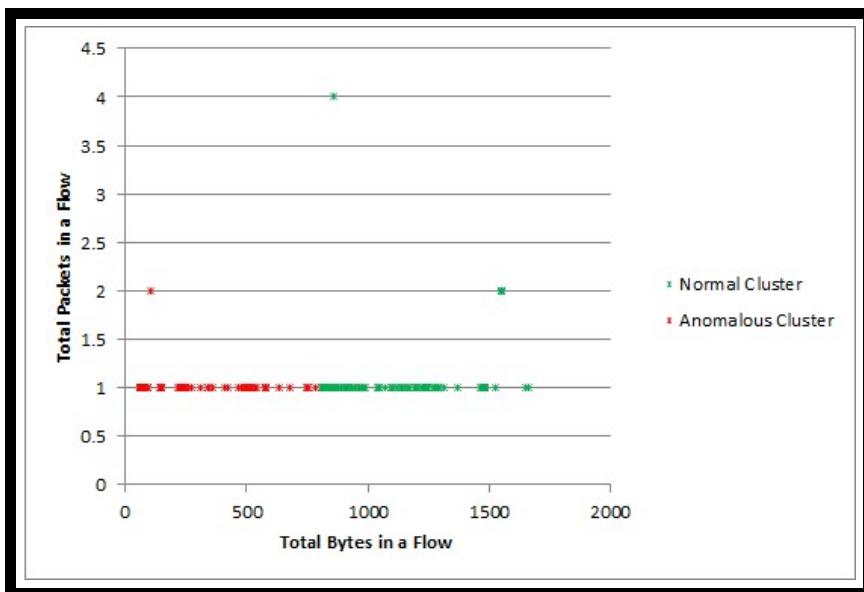


Figure 3: k-means clustering on dataset 3.

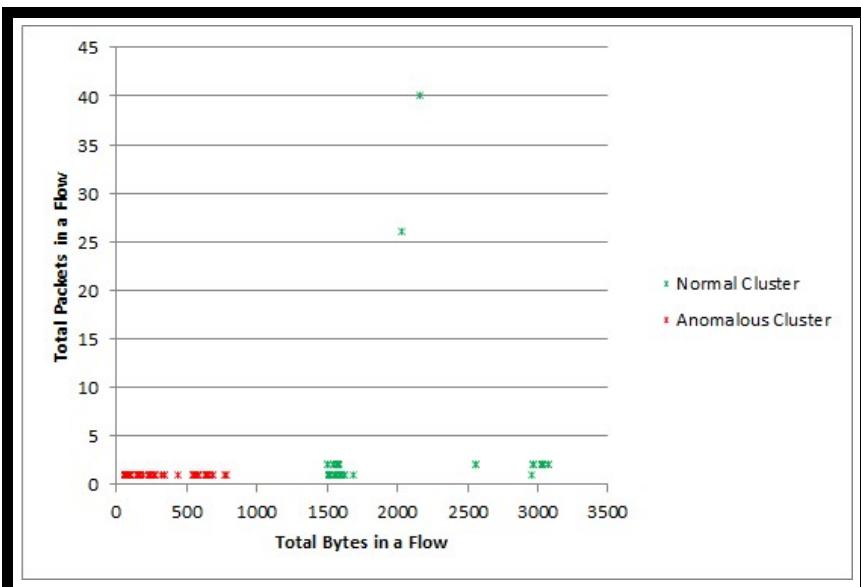


Figure 4: k-means clustering on dataset 4.

The results presented in table 3 shows that for the dataset 1 which consists of mainly TCP injection attacks in the form of several flows having less number of packets and bytes, k-means has assigned 69.35% flows into the abnormal cluster and 30.65% to the normal cluster. In comparison to the traffic characteristics of table 2, the abnormal cluster contains nearly 25% of the normal flows.

The analysis of dataset 2 reveals that k-means have assigned 75.02% flows into the abnormal cluster and 24.98% to the normal cluster. When compared to the baseline traffic characteristics of table 2, the abnormal cluster contains nearly 34% of the normal flows. The analysis of dataset 3 revealed the same characteristics of the k-means algorithm where it has clustered 70.46% flows into the abnormal cluster and 29.54% to the normal cluster. As compared with the traffic characteristics of table 2, the abnormal cluster contains nearly 36% of the normal flows. However, on dataset 4 it was found that k-means was able to cluster 55.55% flows into the abnormal cluster and 44.45% to the normal cluster. k-means has shown improvement in clustering more amount of normal flows into the normal cluster for this dataset. This abnormal cluster contains nearly 11% of the normal flows which is least when compared to other three datasets. This dataset contains highest TCP injection attacks with moderate UDP flooding attacks.

Table 3: Traffic clustering using k-means.

Dataset	Normal Traffic	Abnormal Traffic
Dataset 1	30.65%	69.35%
Dataset 2	24.98%	75.02%
Dataset 3	29.54%	70.46%
Dataset 4	44.45%	55.55%

It could, therefore, be concluded that k-means was able to cluster those TCP flows as abnormal which exhibited the similar type of behaviour in terms of less number of packets and bytes.

Results of Experiment – Outlier detection evaluation

The study analysed the results of outlier detection as shown in table 4 for the input datasets and compared them with the baseline behaviour characteristics results of table 2. For dataset 1, outlier detection was able to classify 28.06% out of 44.69% of the abnormal flows as abnormal. For normal flows, 71.94% in excess to 55.31% were assigned score ≤ 1.2 . In this case, nearly 15% of the normal traffic was given an outlier score greater than 1.2 and were declared as abnormal. For dataset 2, 20.24% out of 41.5% of the abnormal flows were declared as abnormal. For normal flows, 79.76% in excess to 58.5% were assigned score ≤ 1.2 . In this case, again nearly 15% of the normal traffic was given an outlier score greater than 1.2 and were declared as abnormal. The same type of figures exists for dataset 3 where 21.92% out of 46.61% of the abnormal flows were declared as abnormal and 78.08% in excess to 53.39% were assigned score ≤ 1.2 and declared normal. Here also, nearly 15% of the abnormal flows were classified as normal. In dataset 4 which contains many TCP injection traffic, outlier detection has performed badly and classified only 12.07% out of 44.4% of the abnormal traffic as abnormal. The remaining was classified as normal.

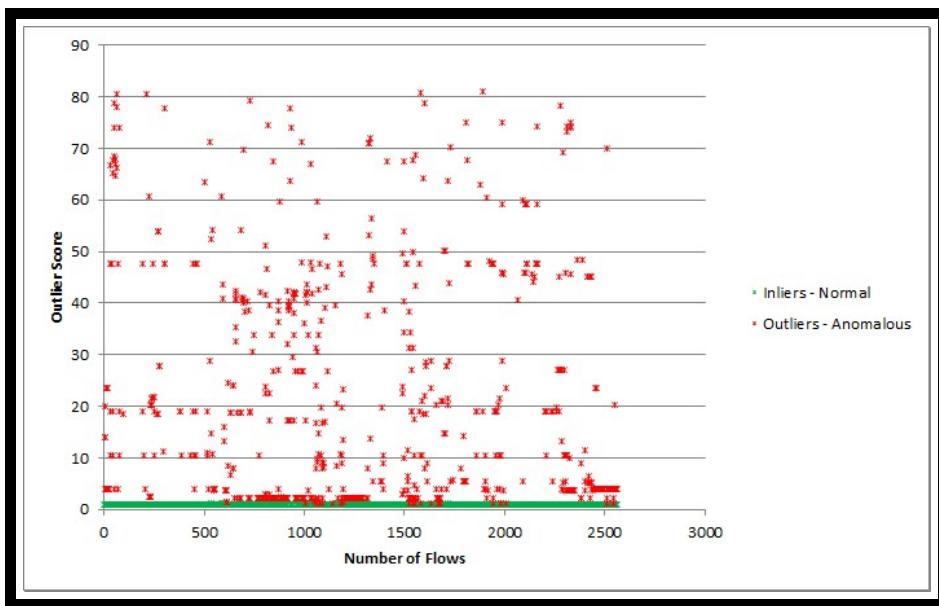


Figure 5: Outliers scores of datasets 1.

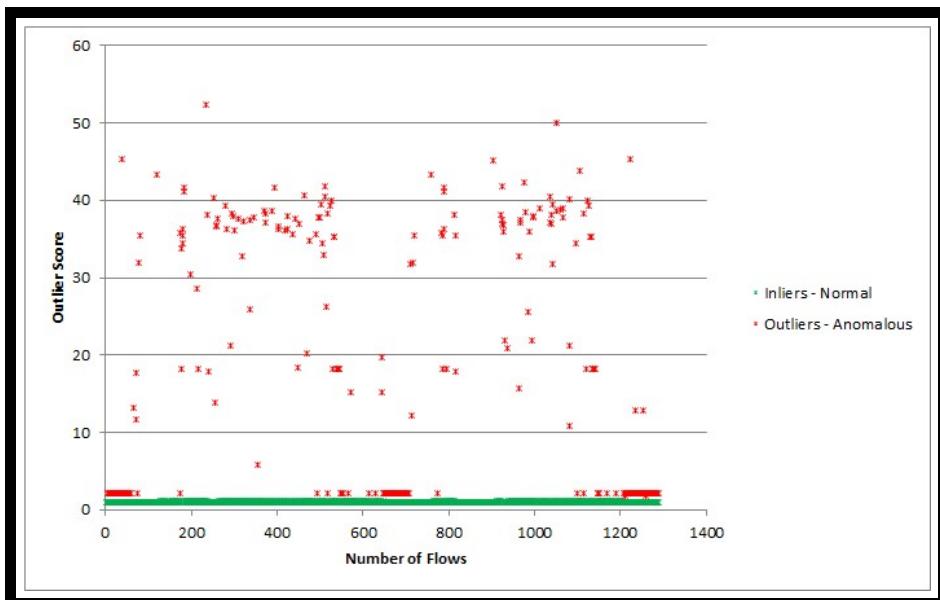


Figure 6: Outliers scores of datasets 2.

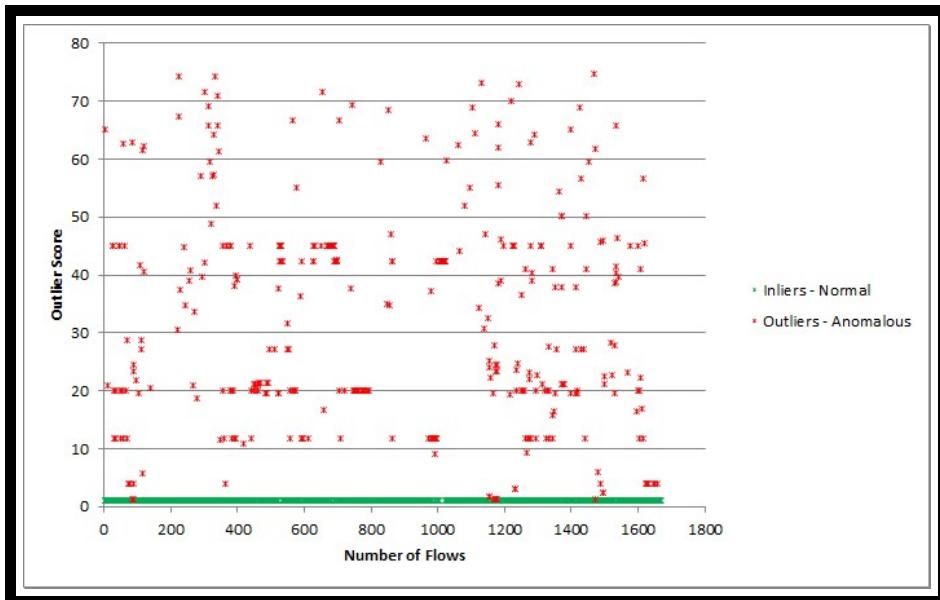


Figure 7: Outliers scores of datasets 3.

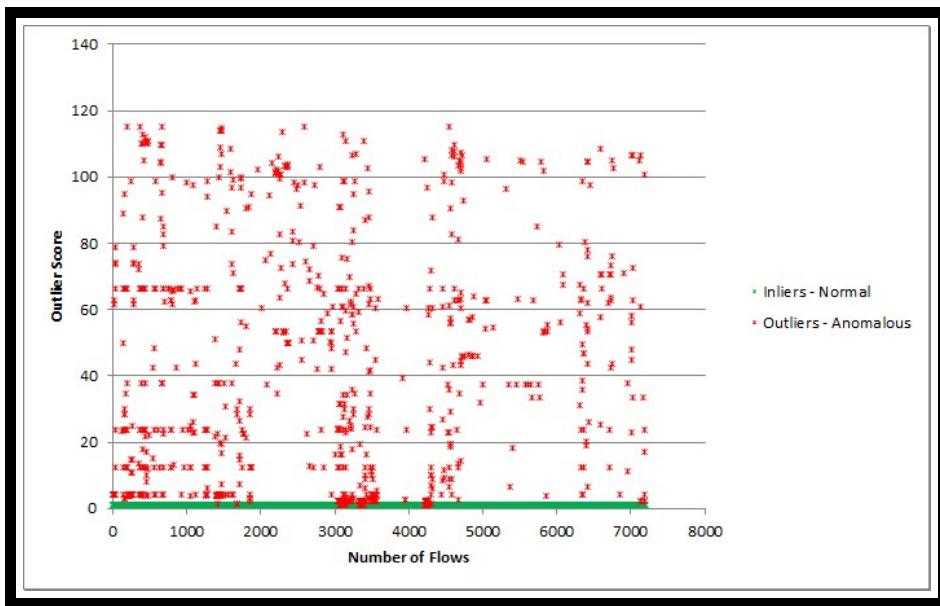


Figure 8: Outliers scores or datasets 4.

Table 4: Traffic classification using outlier detection.

Dataset	Normal Traffic	Abnormal Traffic
Dataset 1	71.94%	28.06%
Dataset 2	79.76%	20.24%
Dataset 3	78.08%	21.92%
Dataset 4	87.93%	12.07%

The following type of abnormal traffic was declared as abnormal by outlier detection.

- ICMP flows originating from different sources to the same destination. (ICMP flooding)
- ICMP flows originating from same sources to the different destination. (ICMP flooding)
- TCP flows originating from different sources and same port to the same destination IP and port and having more than 100 bytes per flow.
- TCP flows originating from same sources to different destination IP and same port and having more than 100 bytes per flow.

The following type of abnormal traffic was declared as normal by outlier detection.

- UDP flows containing more than 500 packets and 10000 bytes per flow. (UDP flooding)
- The following type of normal traffic was declared as abnormal by outlier detection.
- UDP flows originating from different sources to the same destination IP and port.

The study can conclude from the results of outlier detection that outlier detection is better in detecting ICMP flooding. However, it was not able to detect TCP injection where the number of bytes per flow is less than 100 which is generally the case. But in the case of TCP flows where the total bytes exceed 100, it was able to classify them as abnormal which is not always true. However, it was also not able to detect UDP flooding as abnormal in any case.

Performance Evaluation using IDS Metrics

The study evaluated the performance of k-means and outlier detection using performance metrics. The study calculated the metrics for all the datasets for k-means and outlier detection and compared them. These metrics values for both the approaches are shown in table 5. The values for these metrics typically range between 0.0 and 1.0. These values were studied, and the following interpretations were made from these performance metrics.

- **Interpretations for False Positive Rate**

The FPR is the rate with which the IDS categorize normal flows as abnormal flows. According to table 5, it can be seen that for all the datasets, the FPR is higher for k-means and lower for outlier detection. The FPR for outlier detection is half of the FPR for k-means. This is due to the observed facts that k-means has clustered 25-35% of the normal traffic into the abnormal cluster in all the datasets. Thus, it can be interpreted that k-means have high FPR which makes it less effective than outlier detection. In this case, the outlier detection has performed better than k-means. It was also observed that irrespective of the amount and variety of anomalies like TCP injection, UDP flooding and ICMP flooding in the datasets, outlier detection has low false positive rate than k-means.

- **Interpretations for False Negative Rate**

The FNR is the rate with which the IDS categorize abnormal flows as normal flows. A high FNR means that the system is more vulnerable to intrusions. Table 5 shows that k-means have much low FNR than outlier detection for datasets 1, 3 and 4. This means that outlier detection has more tendency to generate no alerts on abnormal flows. The FNR for datasets 1, 3 and 4 is much less in k-means as compared to outlier detection. This means that in the case of TCP injection with less amount of UDP flooding, k-means has less FNR. However, FNR for dataset 2 is slightly more in k-means than outlier detection. Thus, the FNR of k-means in the case of detection of heavy UDP flooding is more or equal to FNR of outlier detection which means that kmeans is slightly more vulnerable to UDP flooding than outlier detection. On an average, outlier detection has high FNR than k-means which make it more vulnerable to anomalies.

Table 5: Performance metrics results.

Metrics	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	k-means	outlier	k-means	outlier	k-means	outlier	k-means	outlier
FPR	0.4	0.2	0.4	0.2	0.4	0.2	0.2	0.1
FNR	0.04	0.6	0.9	0.8	0.01	0.7	0.003	0.8
Sensitivity	0.9	0.3	0.001	0.1	0.9	0.2	0.9	0.1
Specificity	0.5	0.7	0.5	0.7	0.5	0.7	0.7	0.8
CR	0.7	0.5	0.5	0.3	0.7	0.5	0.8	0.5
PR	0.6	0.5	0.4	0.3	0.6	0.5	0.7	0.4

- **Interpretations for Sensitivity**

Sensitivity is also known as the true positive rate (TPR) which is a rate according to which abnormal flows are categorized as abnormal. A more sensitive IDS will also have more FPR. So, there is a trade-off between Sensitivity and FPR. An IDS should not be too much sensitive. If it is too much sensitive, then its FPR is also high. It was observed that k-means is more sensitive than outlier detection for datasets 1, 3 and 4 which contain a high amount of TCP injection. Thus k-means has more tendency to give alerts on abnormal flows as compared to outlier detection in case of TCP injection attacks. As a trade-off, k-means has high FPR than outlier detection. The exception noticed here is for dataset 2 containing the highest amount of UDP flooding where k-means is least sensitive as compared to outlier detection.

- **Interpretations for Specificity**

Specificity also known as the true negative rate (TNR) is the rate with which normal flows are categorized as normal. High specificity means that the IDS is more capable of identifying normal traffic as normal. Table 5 shows that k-means has less specificity than outlier detection. But in the case of large datasets containing more number of TCP injection flows, both k-means and outlier detection has proved to be almost equally capable of identifying normal traffic as normal. Overall on an average, outlier detection is more better in declaring normal flows as normal than k-means.

- **Interpretations for Classification Rate**

The Classification Rate tells the measure of how much accurate the declarations are made by the algorithms. For an IDS, the FPR and FNR should be low with maximum CR. However, it was noticed that k-means is more accurate than outlier detection for all the datasets. Thus k-means is more accurate in identifying more number of TCP injections and UDP flooding flows than outlier detection.

- **Interpretations for Precision Rate**

Precision is also known as the positive predictive value (PPV) which gives the measure of detection of real intrusions in the IDS. More PPV indicates that the algorithm is more capable of detecting abnormal flows. Table 5 shows that k-means is slightly more precise in detecting abnormal flows than outlier detection due to its more PPV. Also, k-means precision increased as compared to outlier detection when the size of the dataset increased.

Performance Evaluation Using Computer Resources

The outcome show that the k-means algorithm consumes 10% to 20% of the CPU and takes approximately 5-10 seconds to execute. On the other hand, outlier detection consumed 50% to 60% of the CPU and takes approximately 40-50 seconds to execute on all the datasets. Figure 9 shows the outcome of this experiment on dataset 1. Similar outcome was noticed for all the datasets and it was found that the outlier detection consumed approximately 5 times more CPU and execution time than k-means.

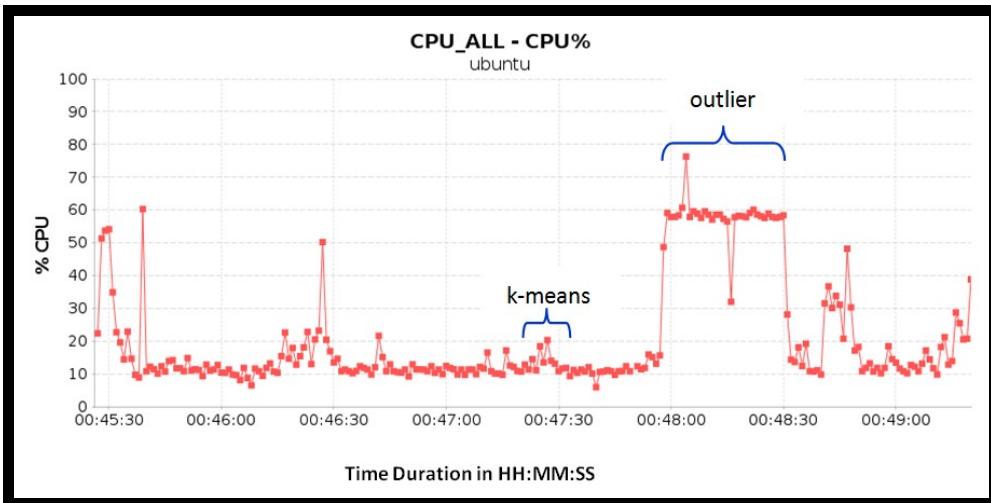


Figure 9: CPU usage on dataset 1.

The study also noticed that the k-means algorithm consumed much less amount of RAM than the outlier detection. Figure 10 shows the amount of RAM freely available during the execution time of the k-means and outlier detection. From this figure, we can infer that around 100 MB of RAM was freely available prior to execution of k-means and outlier detection. The study observed that k-means took an almost negligible amount of RAM as compared to outlier detection which consumed nearly 40% of RAM available freely during its execution. This dropped the free RAM availability to 60 MB. This was observed in all datasets and it was found that the amount of free RAM available during the execution of k-means is 40-45% more than outlier detection.

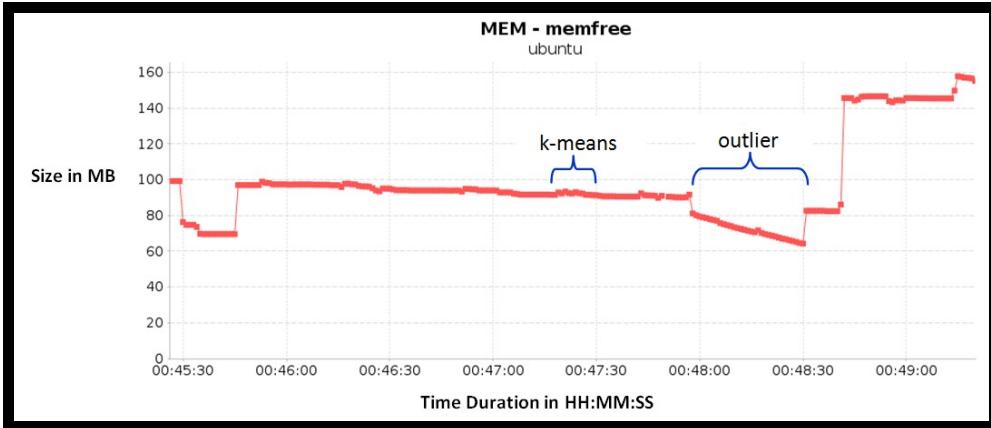


Figure 10: RAM usage on dataset 1.

The study analyzed why outlier detection consumes more CPU and RAM than k-means. The reason for this is that the detection of outliers is significantly involved in the calculation of finding the nearest neighbors of each flow record and that must cross half of the total number of flow records to find the nearest neighbor of each flow record. If n is the total number of flow records, then algorithm for each n must be $n/2$ iterations. This increases the complexity of outlier detection with NOF. Furthermore, this reduces the efficiency of the outlier detection algorithm with respect to k-means. The complexity of both approaches has been solved. For n flow records in a data set, the assignment of each n th flow record to a normal or abnormal cluster can be done

in $O(kn)$ for k-means where k is the number of clusters. However, in the case of outlier detection, finding the nearest neighbors takes $n/2$ iterations for every n th record that brings its complexity to $O(n^2)$.

5. CONCLUSION AND RECOMMENDATION

K-means was able to cluster heavy UDP flooding as abnormal whereas outlier detection has classified it as normal. K-means was better than outlier detection in clustering TCP injections as abnormal containing less number of packets and bytes per flow. But for TCP injections containing more number of packets and bytes per flow, outlier detection performed better. K-means clustered average 65% traffic as abnormal whereas outlier detection classified average 75% traffic as normal.

The study compared the two approaches based on the performance metrics and an amount of computer resources consumed. Outlier detection has low FPR than k-means but proved to be more vulnerable to intrusions than k-means due to the high FNR. However, k-means was more sensitive to outlier detection in generating alerts on abnormal flows. Overall, k-means was more accurate and precise and has better classification rate than outlier detection. Also, the amount of CPU and RAM consumed by outlier detection is much more than k-means which make outlier detection more time expensive.

6.0 FUTURE WORK

The future work includes the optimization of k-means and outlier detection approaches in such a way to detect TCP injections, UDP flooding, ICMP flooding and other DOS attacks with no restriction on the number of packets and bytes per flow. This should be achieved with high classification rate and low false alarm rate. However, the fusion of k-means and outlier detection approaches may help to achieve this, wherein, the k-means would run first to cluster the traffic into normal and abnormal clusters and then calculating the scores for the flows in the abnormal clusters to separate out the normal flows from it. This is because of k-means groups more traffic as abnormal than outlier detection. This may decrease the overall false alarm rate of the IDS and the outlier detection algorithm will also have less number of flow records to traverse through, thereby, consuming less amount of CPU and RAM with fast results. The future work also includes to regularly train the IDS with newer baseline characteristics of normal and abnormal traffic so as to detect newer types of anomalies.

7.0 REFERENCES

- [1] Shah A, Waqas J. Rana, "Performance Analysis of RIP and OSPF in Network Using OPNET", International Journal of Computer Science Issues, Issue 6, No 2, November 2013.
- [2] Lammle, Todd, "CCNA Cisco Certified Network Associate study guide, sixth edition". Indianapolis, Ind.: Wiley. (2007).
- [3] Todorovic Ivana, Sepanovic Stevan, "Measurements of convergence time for RIP And EIGRP Protocols", Scripta Scientiarum Naturalium, volume 2, 2011.
- [4] Deng Justice, Wu Siheng, Sun Kenny, "Comparison of RIP, OSPF and EIGRP Routing Protocols based on OPNET" Final project. Spring, 2014.

- [5] Panford, J. K., Riverson K. & Boansi O. K (2015). Comparative analysis of convergence times between rip, and eigrp routing protocols in a network. Research Journali's; Journal of Computer Science., pg 1-5.

Refactoring to Microservice Architecture

Dr. Latha Sadanandam,

Software Architect,

Danske IT and Support Services India Pvt Ltd.,

Abstract

With the increasing usage of smartphones and other devices, digitization of banking sector is expected to catch up the increasing expectations of the customer. Banks have a significant role in our lives. Every one of us will execute at least a single financial transaction in a day. Hence, it becomes necessary for banks to enrich customer experience. Digitization becomes mandate feature for banks since it is being adopted in all industries in day to day life. Banks love mainframes because only mainframes can provide a single, unified, efficient solution to a host of different problems. Most of the banks uses Mainframe because of its robust, reliable and secured processing power. It also supports the new technologies like mobile, cloud etc., A business case is presented in this paper to explain Micro service and API framework for existing legacy system. Existing architecture is tightly coupled services with less standardization and fair performance. The aim of this paper is to provide solutions to convert the existing architecture to flexible service to support business for time-to-market, increase in performance and operational efficiency and improve customer experience.

1. Application Program Interface (API)

Application Program Interface (API) is used to communicate between programs, data exchange would be part of communication. A common structure has to be agreed between programs for smoother communication. Programs at both ends have to know the location of each other and the structure of the data to be exchanged in advance which is tightly coupled.

A revolution happened by introduction of Service Oriented Architecture framework. In SOA Era, Programs were loosely coupled and data transfer mechanism was done using XML using Web API. Extensible Markup Language (XML) is language and platform independent. This covered the way for producing and exposing APIs over network with better business enablement capabilities including request access, entitlement, identification, authorization, management, monitoring, and analytics. The standard format of data structure (XML Payload) was huge and heavy to be exchanged over network protocol.

A new mechanism was required which is lighter version of XML (SOAP message used in SOA architecture). So Representational State transfer (REST) was introduced. REST is an architectural style of any interface between systems using HTTP to obtain data and generate operations on those data in all possible formats, such as XML and JSON. It has actions to be taken on specific IT resources (File, Image, Database, Service etc.,) using HTTP/HTTPS protocol based on

event triggered. REST verbs used with the protocol are POST, GET, PUT and DELETE.

Since HTTP Protocol is used, URL can be used to access the resources. JSON has quickly become the format of choice for REST APIs. It has a lightweight, readable syntax that can be easily manipulated. The standard request and response to and from API would be in form of JSON.

Benefits of REST API are not limited to

- Improves the portability of the interface to other types of platforms and allows the different components of the development to be evolved independently.
- Increases the scalability of the server application without any disruption since the client and server are loosely coupled.
- Increase in productivity and faster time-to-market.

2. Microservice

Monolithic architecture is traditional server-side systems. The entire system's function is based on a single application. Application can be created with basic features and then scale up over time.

While monolithic architecture puts all of the functions into a single process, Microservice applications break those processes into individual functions. Multiple services are built inside of the application. Each can be tested and developed individually. The services will run as separate processes.

Microservice is a form of Service Oriented Architecture (SOA) style of developing loosely coupled and fine-grained software application as independent deployable services. Each service runs as a unique process and communicates through lightweight protocol to serve a well-defined business function. It allows individual service for continuous enhancement without any disruption and supports continuous delivery and deployment.

3. Refactoring to Microservice

Architecture

Refactoring to Microservice architecture also does not mean to throw away existing application and rewrite new micro services. Rather in majority cases, it is incredible to throw away the application. A methodology has to be identified on how to refactor an existing application using Microservice.

A business capability is a concept from business architecture modelling. It is something that a business does in order to generate value. A business capability often corresponds to a business object. Hence break down monolithic application to services corresponding to business capability. Identifying business capabilities and hence services requires an understanding of the business. This method is stable since the business capabilities are relatively stable.

Once the business capabilities are identified, all the capabilities/Functions cannot be refactored to service. Each service can be built in any technology, since the services are independent from the other and run as separate process. This technology independence enables gradual migration to new modern technologies. But still it is not mandate to replace/migrate the entire application. Legacy code which is robust, secure and highly available can still be maintained in the mainframe system. Mainframe system provides stable solution for security, handling high volume with robust throughput with support of structured data and handling optimized batch workloads to process the data.

Demarcate the business functions as “True core and core surround”. “True core” functions are core capability of the system, for example in core banking system creation of account, Ledger maintenance, interest calculation for accounts etc., can be retained in the legacy system. “Core Surround” functions can be sub activity of the core system such as payment validation, charge

calculation. All these activities can be separated from the core capabilities and migrated to new platform / product.

“True core” functions are exposed as Mainframe services and gradually transform the core functions to the newer technologies. This pattern is termed by Martin Fowler as “Strangler Application Pattern”. This approach is used in transformation strategy of legacy modernization.

4. Microservice Design Patterns

Based on business requirement, functional decomposition is done which provides agility, flexibility and scalability. But the end point of business is to have an application which servers all the functionality. Orchestration of services are required to achieve a business functionality. Design patterns for orchestration are discussed in this section.

- Aggregator Microservice design pattern
- Proxy Microservice design pattern
- Chained Microservice design pattern
- Branch Microservice design pattern
- Shared Microservice design pattern
- Asynchronous messaging Microservice design pattern

4.1 Aggregator Microservice design Pattern

In this pattern, Aggregator act as orchestrator for the service. It might be a simple web page or composite service that invokes multiple services to achieve the functionality required by the application. Since all services are exposed using a lightweight REST services. The web page can retrieve the data and process/display it. If any processing (business logic) is required for the data retrieved from services, computing program (bean) can be developed that would apply the business logic and computation on the data to be displayed on the web page.

An advantage of this pattern is that the individual services can evolve independently and the business need is still provided by the composite microservice.

4.2 Proxy Microservice design pattern

In this case, no aggregation might be required, but a different microservice can be invoked based upon the business need. The proxy may be a *dumb proxy* in which case it just delegates the request to one of the services.

Alternatively, it may be a *smart proxy* where some data transformation is applied before the response is served to the client. An example of this would be where the presentation layer to different devices can be encapsulated at the proxy level.

4.3 Chained Microservice design pattern

Chained microservice design pattern produce a single consolidated response to the request. The key feature is that the client is blocked until the complete chain of request/response. The synchronous nature of the chain will make client to wait for long time.

Advantage of this pattern is any service can be removed or added as the per the business requirement.

4.4 Branch Microservice design pattern

Branch microservice design pattern extends Aggregator design pattern and allows simultaneous response processing from multiple chains of microservices. Service either a web page or a composite microservice, can invoke two different chains concurrently in which case this will resemble the Aggregator design pattern. Alternatively, Service can invoke only one chain based upon the request received from the client.

4.5 Shared Microservice design pattern

In this design pattern, some microservices might share caching and database stores. This would only make sense if there is a strong coupling between the two services. This might be a business needs to use either the database or caching.

This act as anti-pattern of microservice because of tight coupling and sharing of resources across services.

4.6 Asynchronous messaging Microservice design pattern

A blend of REST call and pub/sub messaging may be used to realize the business need in this pattern. In this design pattern, Service may call a service synchronously which might then communicate with Service asynchronously using a shared message queue.

This pattern has the limitation of being synchronous and the response time is slow.

5. Architecture framework for exposing Legacy application as REST services

Most often the demarcated “True core” functions resides on Mainframe. The best way is to expose these applications as REST services/API in digital era for the customer touch points like Mobile etc., to maximize the value of the system.

There are multiple models to perform REST on mainframe. Let us discuss here on three architectural model.

- Exposing REST calls at Mid-layer (Data power Appliance)
- Exposing REST calls in CICS by exposing applications as JSON web service
- Handling REST calls using Gateway

Let us discuss on the architecture of these three models with its pros and cons.

5.1 Exposing REST calls at Mid-layer

In this model, the REST API call can be handled in a mid-layer device such as Datapower or IIB. The rest resource, CICS web service (non-REST) is processed in Mainframe system and the response is provided for the request. The protocol could be Messaging protocols (MQ, JMS etc.). It has less changes on mainframe legacy system. But the drawback is that mapping between legacy system data structure and requester data structure has to be handled in mid-layer manually by using wrappers.

This architecture model uses Asynchronous messaging Microservice design pattern which has its own limitation of response latency and not being synchronous.

5.2 Exposing REST calls in CICS by exposing applications as JSON web service

REST call can be made directly to CICS. CICS application has to be exposed as JSON Web service, configuring and deploying the services in CICS transaction region. A proxy is required to access the services in the CICS region of legacy system.

CICS JSON Web service can be achieved using two different approaches – Top-down and Bottom-up.

In Top-down approach, using JSON schema from client is used to create copybook using DFHJS2LS utility. The copybook can be used in the wrapper program to map the structure to CICS source program.

In Bottom-Up approach, JSON web service are created using request and response copybooks using DFHLS2JS utility.

This pattern is a direct way of invoking CICS services, but requires different methods for each type of CICS application.

Proxy microservice design pattern is applied to this architecture. In this case, no aggregation might be required, but a different microservice can be invoked based upon the business need. Proxy help in data transformation before the response is served to the client. The response can accommodate request from any type of client. The encapsulation is done at the proxy layer.

5.3 Handling REST calls using Gateway

Using Gateway in LPARs to handle the REST API calls. The gateways provide controllable entry point for LPARs (e.g., Z/OS connect). Z/OS connect Enterprise Edition runs on liberty profile on zos. It act as receiving server for URL, parse the URL and identify the resource. Z/OS connect EE is configured such that identified resources is mapped to a CICS program on a specific region.

But CICS handles the I/O structure in COMMAREA/CHANNELS copybook structure, Z/OS connect EE has provision to convert JSON request to COMMAREA/CHANNELS copybook structure and vice versa.

Execute the CICS program based on the query parameter and return back the output copy book structure back to Z/OS connectEE which converts in JSON Payload.

This pattern is standard approach, since auditing and tracing is easy and controls the call based on security criteria.

Aggregator or chain or branch microservice design pattern is applied in this framework. This patterns helps the model to adopt a solution, that the individual services can evolve independently and the business need is provided by the composite microservice.

6. Conclusion

A microservices architecture breaks down application components into small, manageable services, each running within its own function independently by a different team. This makes it easier to change functions and qualities of the system at any time. Microservices lends itself to continuous delivery software development, because a change to a small part of the application only requires one or a small number of services to be redeployed. The services can be designed and developed as completely autonomous entities, with individual services for new applications being reused or integrated with third-party services without affecting other applications.

With microservices-based architecture, implementation can be done using any technologies and frameworks. Each microservice can be built by any set of languages or tools, since each is independent from the others and each runs a separate process. This technology independence also means that individual services within an application can be gradually replaced with applications based on more modern technologies—without having to replace the entire application. Microservices are independent and agile with rapid application evolution. Organizations can respond more quickly to customer and market feedback, and releases no longer need to be delayed by the schedule of a single release.

Acknowledgement

The author would like to thank the **Management team, Danske IT and support services India Pvt Ltd.**, for providing support and providing permission to publish the paper.

The author would also like to acknowledge and thank the authors and publishers of referenced papers for making available their invaluable work which served as excellent input for this paper.

BREAST CANCER STAGE CLASSIFICATION ON DIGITAL MAMMOGRAM IMAGES

Dr. G. Rasitha Banu¹, Fathima N Sakeena², Mrs.Mumtaj³, Mr.Agha Sheraz Hanif⁴

¹Assistant Professor, Faculty of PHTM,Dept. of HI,Jazan University,KSA

² Lecturer,Faculty of CS& IS,Jazan University,KSA

³ Assistant Professor,Dept.of bioinformatics,Md.Sathak college,India

⁴ Lecturer,Faculty of PHTM,Department of HI,Jazan University,KSA

ABSTRACT

Breast cancer is a disease in which the cells of the breast grow out of control, creates an abnormality in the breast tissue. It is the second leading cause of death in women worldwide. In Saudi Arabia, Ministry of health reported that the number of new cases of cancer is 2741 including about 19.9% of breast cancer in women due to unawareness , it usually occurs in women at the age of 52. It accounts for about 22% of all new cancers in women. In developing countries there are still large numbers of breast cancers diagnosed in later stages. So the death rate is also high. To prevent people from this disease, it should be detected at an earlier stage which reduces death rate. Digital mammogram is used for this purpose. The suspected symptoms causing breast cancer are age, post menopause, stress, family history, physical inactivity, obesity, hormonal imbalances and genetically mutated abnormalities. Our work focus on detecting stage of breast cancer using image processing techniques and data mining technique is used to classify the stage of breast cancer and the performance of classifier is evaluated through confusion matrix.

Key Words: Image Processing,Data mining,Weka,Classification,J48.REPTree

1. Introduction

Breast cancer stage is described the condition of cancer, based on its location, its size, where it spreads and the extent of its influence on other organs. In general, the level of breast cancer varies from stage 0 to stage IV. Among various diagnostic techniques, such as X-ray, MRI, breast ultrasound, digital mammograms are the most reliable and inexpensive to detect the symptoms of breast cancer at the early stage, can disclose many information about these abnormalities like masses, micro calcifications, architectural distortion and bilateral asymmetry.

Digital Mammogram is one of the efficient technique to detect the cancer at an earlier stage. There is a special detector which converts a X-ray energy into digital image. It helps the people to reduce the mortality rate. It detects abnormalities easily. It is advisable to all women should do regular screening test in the age of 35 to prevent from this disease. There are many advantages of digital mammogram such as: patient spend less time for screening, radiologist quickly transmit the images to another physician and they can be easily manipulated.

Data Mining is a process of discovering hidden patterns in the database. There are many techniques available such as neural networks, association rule mining, classification and clustering and so on. In our work, we have used data mining tool weka to classify the stage of breast cancer from digital mammographic images.

2. Objective:

1. The main objective of our work is to detect the stage of breast cancer from digital mammographic images based on area of size of the pixel.
2. This computer aided diagnostic system is used is support the radiologist to determine the stage of breast cancer and as an aid in decision making.
3. Classifying the stage of the breast cancer using data mining classification techniques.

3. Proposed Methodology:

Breast masses and micro calcifications are the main indications of abnormalities in digital mammograms. Breast cancer detection can be carried out by using various image processing techniques. The proposed method involves data collection, image preprocessing, segmentation of ROI, feature selection and classification of cancer stages in abnormal mammograms.

1. **Data collection:** Mammography Image Analysis Society (MIAS) database used in this research. Data is in the form of PGM (Portable Gray Map) format. In this research, 50 mammogram images are used for determining the various stages.
2. **Preprocessing:** The noise removal is done by using Gaussian filter. Gaussian smoothing is very effective for removing Gaussian noise, the degree of smoothing is controlled by σ , which is set as 1. The contrast of mammogram image is increased by using Cumulative Histogram Equalization, which has good performance.

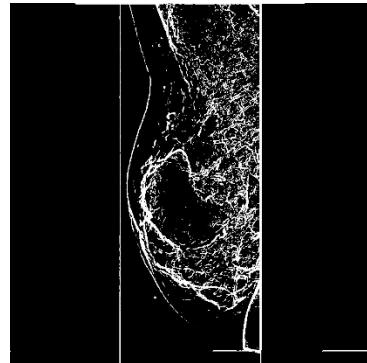


Image1: After preprocessing

3. **Segmentation of ROI:** Segmentation is the process of partitioning a digital image into multiple segments. Segmentation can be carried out by using local thresholding. Edge detection is used to divide into areas corresponding to different objects to enhance the tumor area in mammographic images.
4. **Feature extraction and selection:** Using ROI, the area of size of the pixel can be calculated to identify the various stages of the breast cancer.
5. **Classification:** The process of assigning a label to unknown objects. It is a supervised learning, the image attributes (features) are given as the input to data mining classifiers such as J48 and RepTree to classify the stage of the breast cancer on digital mammograms.

4. Experiments with Weka:

In this research, 50 malignant mammogram images from MIAS database are used, where 16 images are from the group of malignant mammogram dense-glandular, 16 malignant mammogram images derived from fatty group and 18 malignant mammogram images derived from fatty-glandular groups. After the process of preprocessing, segmentation tumor area will be identified. Further using Region of interest the area of pixel can be calculated. Depends on the value of the pixel, the stage of the cancer to be identified. The following table shows Table 1 show the result of determining the stage of cancer from malignant digital mammogram images.

Ref No	Tissue	abnormality	Severity	Radius	area	class
mdb023	G	CIRC	M	29	22268	1
mdb028	F	CIRC	M	56	9385	1
mdb058	D	MISC	M	27	8698	1
mdb072	G	ASYM	M	28	22342	1
mdb075	F	ASYM	M	23	11328	1
mdb090	G	ASYM	M	49	39032	2
mdb092	F	ASYM	M	43	5184	1
mdb095	F	ASYM	M	29	34833	2
mdb102	D	ASYM	M	38	30786	2
mdb105	D	ASYM	M	98	161097	4
mdb110	D	ASYM	M	51	45413	2
mdb111	D	ASYM	M	107	56732	2
mdb115	G	ARCH	M	117	81616	3
mdb117	G	ARCH	M	84	47906	2
mdb120	G	ARCH	M	79	67896	3
mdb124	G	ARCH	M	33	26426	2
mdb125	D	ARCH	M	60	31840	2
mdb130	D	ARCH	M	28	74694	3
mdb134	F	MISC	M	49	6505	1
mdb141	F	CIRC	M	29	63602	3
mdb144	F	MISC	M	27	20944	1
mdb155	F	ARCH	M	95	6957	1
mdb158	F	ARCH	M	88	641	0
mdb170	D	ARCH	M	82	11499	1
mdb171	D	ARCH	M	62	162560	4
mdb178	G	SPIC	M	70	13680	1
mdb179	D	SPIC	M	67	65330	3
mdb181	G	SPIC	M	54	24702	1
mdb184	F	SPIC	M	114	32590	2
mdb186	G	SPIC	M	47	2535	0
mdb202	D	SPIC	M	37	1901	0
mdb206	F	SPIC	M	17	12891	1

mdb209	G	CALC	M	87	57756	2
mdb211	G	CALC	M	13	9913	1
mdb213	G	CALC	M	45	5656	1
mdb231	F	CALC	M	44	39429	2
mdb238	F	CALC	M	17	186754	4
mdb239	D	CALC	M	25	156879	4
mdb241	D	CALC	M	38	37691	2
mdb249	D	CALC	M	64	1426	0
mdb253	D	CALC	M	28	58355	2
mdb256	F	CALC	M	37	9141	1
mdb264	G	MISC	M	36	32455	2
mdb265	G	MISC	M	60	66420	3
mdb267	F	MISC	M	56	41947	2
mdb270	G	CIRC	M	72	9738	1
mdb271	F	MISC	M	68	1949	0
mdb274	F	MISC	M	123	11251	1
mdb245	F	CALC	M	38	10734	1
mdb250	D	CALC	M	64	2956	0

Table 1. : The results of determining stage of cancer from Digital Mammogram image.

Out of 50 images, 6 images are belong to stage 0, 19 images are belong to stage I, 15 images come under stage II, 6 images come under stage III and 4 images are belong to stage IV.

The open source software Waikato Environment for knowledge Analysis 3.7(WEKA) is used for our experiment. It is a collection of machine learning algorithms for data mining tasks. Weka can be downloaded from the website ¹⁰.

4.1 Performance Measure of Classifiers:

In our experiment, breast cancer data is supplied to classifier of J48, and Random tree algorithms to classify the stages of breast cancer. The classifiers performance are evaluated through Confusion Matrix.

a. Confusion Matrix

It is used for measuring the performance of classifiers. In the confusion matrix, correctly classified instances are calculated by sum of diagonal elements TP (True Positive) and TN (True Negative)

and others as well as FP (false positive) and FN (False Negative) are called incorrectly classified instances.

b. Accuracy

It is defined as the ratio of correctly classified instances to total number of instances in the dataset.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

5. Result Analysis:

There are totally 50 records in the breast cancer dataset. Among these 19 instances belongs to stage 0, 15 instances belongs to stage I, 4 instances belongs to stage II, 6 instances belongs to stage III, 6 instances belongs to stage IV. The following table shows confusion matrix with 12 attributes.

The following Table 2 represents confusion matrix for Random Tree Algorithm

Target class	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	18	0	0	1	0
Stage I	9	3	0	3	0
Stage II	2	2	0	0	0
Stage III	3	2	0	1	0
Stage IV	5	1	0	0	0

Table 2: Confusion matrix for Random Tree Algorithm

In Random tree classifier, the correctly identified instances are 22 and incorrectly identified instances are 28.

The following Table 3 represents confusion matrix for J48Algorithm.

Target class	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	18	1	0	0	0
Stage I	0	14	0	1	0
Stage II	0	0	4	0	0
Stage III	0	0	1	5	0
Stage IV	1	0	0	0	5

Table 3: Confusion matrix for J48 Algorithm

In J48 classifier, the correctly identified instances are 46 and incorrectly identified instances are 4.

The following Table 4 depicts detailed accuracy of J48, Random Tree algorithm

Classifier	Accuracy
Random Tree	55.55%
J48	96.66%

Table 4: Accuracy of classifiers

Table 4 shows that J48 is giving highest accuracy.

The following chart1 shows the accuracy of classifiers.



Chart 1: Performance Analysis of classifiers

In this chart, X axis represent the algorithm and Y axis represent the accuracy. It shows that the accuracy of J48 is 96.66 % which is best than Random Tree Algorithms.

6. Conclusion

In our research, 50 mammogram images from MIAS database are used. We have used image processing techniques such as Gaussian filtering, histogram equalization, thresholding, edge detection are used to remove the noise, enhance the image, and find the region of interest. The image attributes are extracted from the processed image, according to the area of the size of the pixel, stage of the breast cancer identified. Out of 50 images, 6 images are belong to stage 0, 19 images are belong to stage I, 15 images come under stage II, 6 images come under stage III and 4 images are belong to stage IV. The breast cancer stages are classified using data mining classifier such as J48 and Rep Tree. The performance of the classifiers are evaluated though confusion matrix in terms of accuracy, in which J48 provides good accuracy.

7. References

1. Karmilasari et.al "Sample K-Means Clustering Method for Determining the Stage of Breast Cancer Malignancy Based on Cancer Size on Mammogram Image Basis", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 3, 2014
2. Sezgin, M., and Sankur, B., "Survey over image thresholding techniques and quantitative performance evaluation". Journal of Electronic Imaging, 13 (1): 146–165. 2004.
3. Maitra, I.K., Nag S., Bandyopadhyay S.K., "A Novel Edge Detection Algorithm for Digital Mammogram", International Journal of Information and Communication Technology Research, Vol 2 No.2, February 2012.
4. Kamdi, S., "Image Segmentation and Region Growing Algorithm", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Vol 2 Issue no.1, October 2011.
5. Priya, D.S., and Sarojini, B., "Breast Cancer Detection In Mammogram Images Using Region-Growing And Contour Based Segmentation Techniques", International Journal of Computer & Organization Trends, Vol.3 Issue 8, September 2013.
6. Martins, L.d.O., Junior, G.B., Silva, A.C., Paiva, A.C. and Gattass, M., "Detection of Masses in Digital Mammograms using K-Means and Support Vector Machine", Electronic Letters on Computer Vision and Image Analysis 8(2) : 39-50, 2009.
7. S.P. Meharunnisa et.al "Detection of masses in digital mammograms using SVM", IJCTA, Vol 8 Issue no.3,2015,ppno:899-906.

OPTIMIZING BUILDING PLAN FOR A (9m X 12m) HOUSE USING LEARNING SYSTEMS

Dr. Khalid Nazim S. A.¹, Dr. Harsha S², Abhilash Kashyap B³, Dr. Fayed Al Fayed⁴

1. Assistant Professor, Department of CSI, College of Science, Majmaah University, Majmaah 11952, Saudi Arabia, k.sattar@mu.edu.sa,
2. Associate Professor, Department of ISE, JIT, Bengaluru, harsha.s@jyothiyit.ac.in
3. 6th CSE, JIT, Bengaluru, abhilashb.96@gmail.com, nagarjun@protonmail.com, sandeepbadrinarayan@gmail.com
4. Assistant Professor, Department of CSI, College of Science, Majmaah University, Majmaah 11952, Saudi Arabia, falfayed@mu.edu.sa

Abstract- Computer aided design (CAD) has increased by orders of magnitude the power of design tools available to the engineer. Advantages of CAD include the reduction of computation time and therefore its cost, the elimination of the amount of tedious and error-prone detailed calculations done by the engineer, and the ability to develop and analyze much more complete models of structures. All present applications of the computer to structural design deal with later stages of the design process, namely, analysis, proportioning and drafting. In Architecture more prominence is given to outlook and not aesthetics and we as engineers should consider this as a problem and give a solution in the form of optimization. With the advancements of a section of computer science called artificial intelligence, it is now conceivable to create a knowledge-based system to automate or assist in the early, preliminary stages of the civil engineering design process. The purpose of this work is to try and design a set of algorithms to solve the building design problem as an optimization issue.

Keywords: Artificial intelligence, Design process, Knowledge based system, Optimization, Site Management & Buildings department(SMB).

A. Introduction

"Artificial intelligence is the study of ideas which enable computers to do the things that make people seem intelligent"[2]. Ideas are being developed to facilitate the creation of knowledge-based systems using the experience and knowledge of experts. The civil engineering problems are not repetitive, as the problem definition is always influenced by several factors like financial modes, importance of structure and site conditions and so on. Therefore, although the use of computers in structural analysis started almost four decades ago, the profession has not been able to make use of computers fully, especially, for structural design and planning. This is mainly because of problem specific nature, need for logical reasoning, feasibility constraints and use of experience required in actual design process and planning. Expert systems have capabilities to incorporate some of these requirements for programming a machine for solving a design problem and algorithms are usually constructed with the natural counterpart in mind. Algorithms are intended to solve problems with extreme objectivity. Each algorithm is designed to solve a specific problem with a crisp set of variables. Genetic algorithms on the other hand deal with fuzzy sets where a range of variable values are existing.

B. LITERATURE REVIEW

According to <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1019&context=cee>, they quote

"In the present practice of preliminary design very little optimization is done and selection is based on rough calculations. Computer assisted preliminary design would provide opportunity for optimization by consideration of a much larger range of alternatives"

C. DESIGN PROCESS

Design can be viewed as the general process in which an idea is developed into detailed instructions for manufacturing a physical product. The design process starts with a definition of a need. The activities that follow can be grouped into four phases [2]:

1. *Synthesis*: The clarification of the input parameters and their interaction to create a structure that will meet design requirements.
2. *Analysis*: The modelling and solving of equations to predict the response of a selected structure.
3. *Evaluation*: The activity of placing a worth on the structure where worth may be cost, safety, or energy consumption.
4. *Optimization*: the search over the range of possibilities to improve the design as much as possible.

Preliminary design is part of the synthesis phase. In preliminary design alternative structures are developed, a structural configuration is selected, and the preliminary parameters of components are determined.

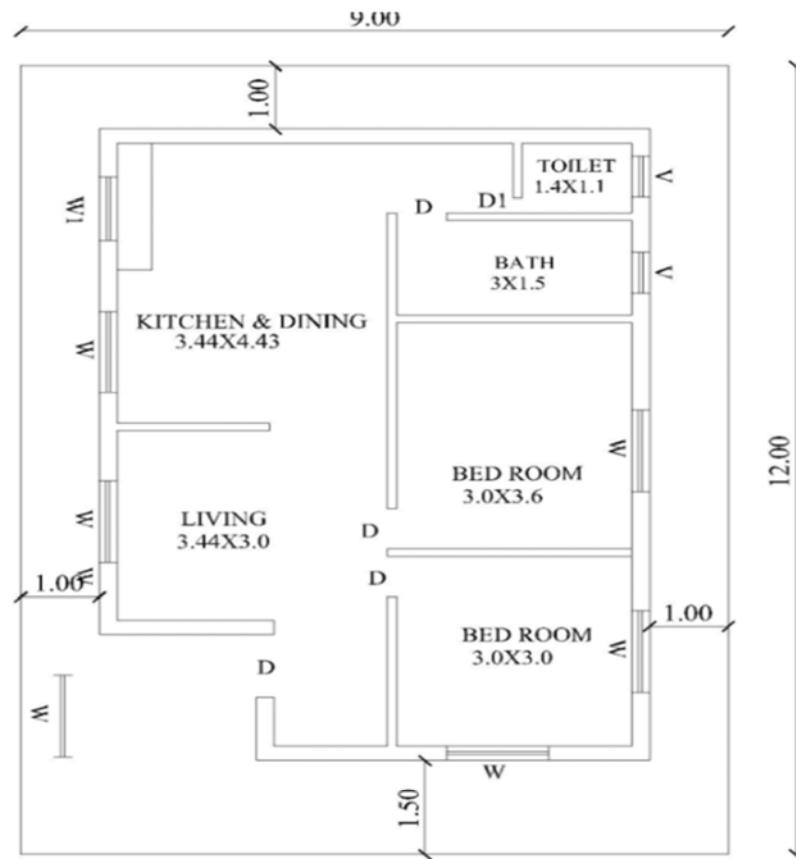
In the present practice of preliminary design very little optimization is done and selection is based on rough calculations. Computer assisted preliminary design would provide opportunity for optimization by consideration of a much larger range of alternatives. It is possible that computer assisted preliminary design could free the engineer from the implementation of existing structural schemes and allows him to pursue new schemes.

To begin a discussion of CAD, a distinction will be made between computer aided design and design automation [3]. In computer aided design man and machine work together on a problem using the best characteristics of each. In design automation the computer deals with all the demands and constraints without recourse to the designer. The latter may be suitable for design of components, but the method is inherently inflexible, and the exclusion of the designer often leads to dissatisfaction. There are two schools of thought regarding the consequences of CAD. For one, computers remove the repetitive tasks and make room for creativity. The opposing view is that computers stifle creativity by distancing the designer from design. An optimum CAD strategy would be to remove the repetitive tasks without creating a large gap between the designer and the design process.

Today, CAD in structural engineering involves almost exclusively analysis, proportioning of structural components, and production of drawings and schedules. There are very few applications to conceptual and preliminary design. Conceptual and preliminary design are considered the creative aspects of design. Yet, generally the preliminary design process is not new design but redesign, where redesign involves the application of existing structural ideas to a design. New design implies the development of a new structural configuration. Redesign is the application of a set of rules to assign values to predefined variables. Thus, it appears that preliminary structural design process may be placed in a knowledge-based system, where IF THEN rules are used to instantiate values in a data structure [4].

A knowledge-based program is developed using the knowledge of experts. Once the program is developed there should be close interaction between the designer and the computer [5]. The computer should be able to respond to queries on the design process as well as accept additional information. Since a design prepared by the computer follows a limited number of rules, close supervision by the designer is necessary. In this way the designer will realize inadequacies in the existing set of rules and make revisions or additions to the rules when necessary.

D. METHODOLOGY



GROUND FLOOR PLAN

SCHEDULE OF OPENINGS:

DOORS

D	0.90X2.10
D1	0.75X2.10

WINDOWS

W	1.20X1.37
W1	1.00X1.37

VENTILATOR

V	0.60X0.60
---	-----------

Fig. 1. The schedule of openings for the ground floor plan

E. PSEUDO CODES

SPACE UTILIZATION

If the building is enclosed only with outer walls
Then space utilization = 100
Else If its provided with interior partition walls which does not lead to passages
Then space utilization = 75%
Else if its provided with interior partition walls which lead to passages
Then space utilization = 50%
Else If the building takes the shape of a maze
Then space utilization = 0

UTILIZATION OF DAYLIGHT

If the window type is fixed with glass panes and completely closed
Then intensity = 70%
If the window type is fixed with glass panes completely closed with opaque curtains
Then intensity = 0
If the window type is fixed with glass panes having translucent type of curtains
Then intensity = 60%
If the window type is fixed with glass panes having transparent type of curtains
Then intensity = 70%
If the window type is fixed with tinted glass
Then intensity = 50%
If all the above conditions are provided with mesh
Then intensity is reduced by 20%
If the doors provided in the periphery of the building are completely open
Then intensity = 80%
If the doors provided inside the building
Then intensity = 60%
If the doors are partially open
Then intensity = 40%
End if

ENERGY

$l_1=9.68, b_1=7.13$
For ($l=1; l \leq l_1; l++$)
For ($b=1; b \leq b_1; b++$)
Solar Intensity=125
Energy= $l * b * \text{Solar Intensity}$
If (energy ≥ 5400) then
output ($l * b$)
Exit for

WATER CONSUMPTION

Annual water requirement for 5 people = 547500 lt
If maximum catchment area of 111 m² is utilized
Then annual water harvesting potential = 64,646.4 lt
If average catchment area of 55 m² is utilized
Then annual water harvesting potential = 32,323.2 lt
Else optimum water harvesting cannot be achieved
End if

MATERIALS

'M' main walls and 'N' partition walls

If (M=12 & P=8)

Then quantity of earth work excavated = 22.85 m³

If the SMB is (0.24*0.11*0.07) m³

Then number of SMB = 12955

End if

F. ANALYSIS

Optimization is achieved using the algorithms mentioned with feedback from user as well. Once the design is complete we compare the design with conventional designs for the utility parameters. This section discusses the results and analysis of the comparison.

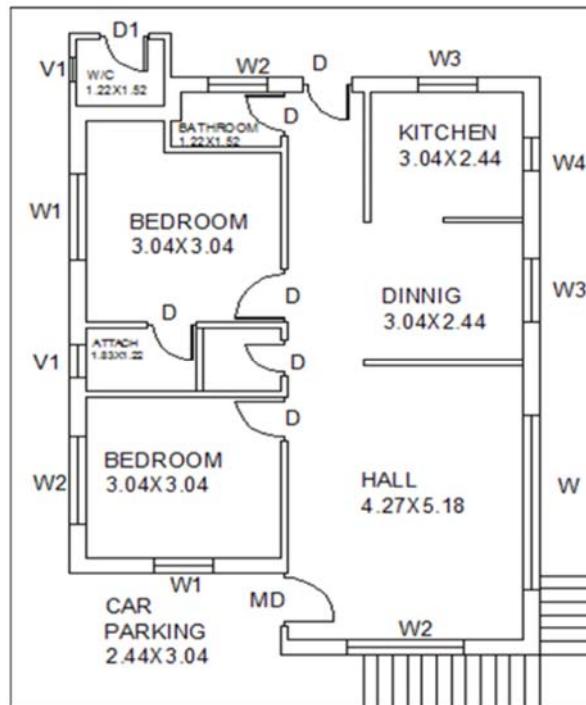
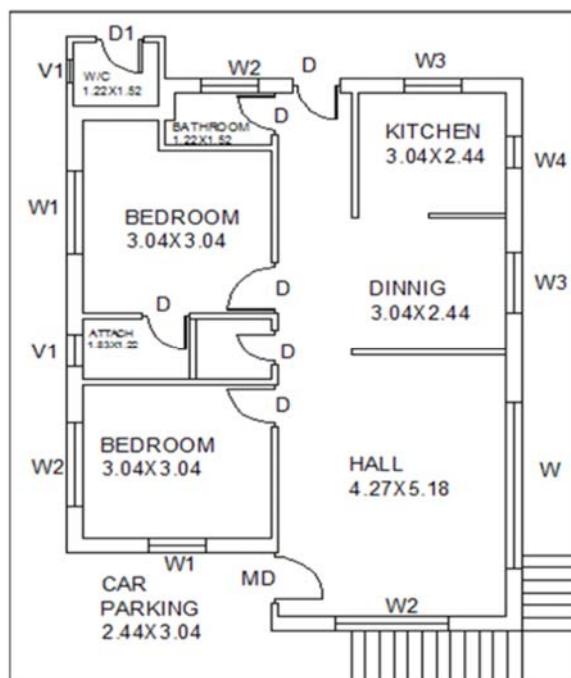


Fig. 2. The design showing the conventional values

TABLE I. Depiction of dimensions of the house plan optimized

NOTES	
MD 1.22 X 2.13	W 1.83 X 1.83
D 0.91 X 2.13	W 1 0.91 X 1.83
D1 0.91 X 2.13	W2 1.52 X 1.25
V 0.91 X 0.61	W3 0.94 X 1.2
V1 0.35 X 0.61	W4 0.35 X 0.61

G. OPTIMIZED PLAN



SCHEDULE OF OPENINGS:

DOORS	
D	0.90X2.10
D1	0.75X2.10
WINDOWS	
W	1.20X1.37
W1	1.00X1.37
VENTILATOR	
V	0.60X0.60

Fig. 3. The plan showing the optimized values

TABLE II. Comparison of conventional and optimized parameters

PARAMETER	CONVENTIONAL (%)	OPTIMIZED (%)
1.Space Utilization	40-55	60-65
2.Utilization of daylight	variable	70
3.Energy	variable	100
4.Rain water harvesting potential	Not considered	60
5.Materials(SMB)	Not considered	100

H. CONCLUSIONS

Optimization for civil engineering plans in architecture is a relatively a new area that requires a lot of work. In this paper, we have attempted to combine fuzzy logic [6] with algorithm and the results are promising. Optimized designs will not only increase the effectiveness of construction but also improve living conditions for people. Such designs will very soon become necessary in this ever-changing economy, increasing impact of population growth on environment, for effective utilization of available resources and preservation of nature.

Using these algorithms and different AI languages an expert system can be developed for the domain to solve several civil engineering problems such as for analysis-design, concrete technology [7], design of R.C.C. and structural steel components behavior modeling of fiber reinforced concrete beams and predicting large deflection response of rectangular plates [8] which can be taken as the further scope of study. Also, these algorithms are preliminary steps which need further refinement to build a toolkit that can efficiently design a house of any dimension and any number of floors. Nevertheless, the platform is still in its infancy and continued growth and development in the field is certainly evident from the recent trends.

REFERENCES

- [1] <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1019&context=cee>
- [2] Journal on application of artificial intelligence on structural engineering A.K.L SRIVASTAV, The Experiment, OCTOBER 2012, Vol. 3(3), 199-202
- [3] International journal on artificial intelligence tool Vol. 15, No. 6 (2006) 867-874 © World Scientific Publishing Company
- [4] Artificial Intelligence and Expert Systems, Dan W. Patterson, PHI learning private limited, New Delhi
- [5] Artificial Intelligence, Elaine Rich, Kevin Knight, TATA McGraw HILL Edition
- [6] "Fuzzy sets". Information and Control. Zadeh, L.A. (1965), 8 (3): 338–353. doi:10.1016/s0019-9958(65)90241-x
- [7] Cost Estimation of Structures in Commercial Buildings, 1991, Surinder Singh, Google books
- [8] Large deflections of rectangular plates, K.T. Sundara Rajalayengar, M.MatinNaqvi, DOI: 10.1016/0020-7462(66)90024-2, International Journal of Non-Linear Mechanics, Volume 1, Issue 2, October 1966, Pages 109-122

A Survey on Rule-Based Systems and the significance of Fault Tolerance for High-Performance Computing

G. Sreenivasulu

Department of CSE, J B Institute of
Engineering and Technology,
Hyderabad, India
svgonuguntla@gmail.com

P. V. S. Srinivas

Sreenidhi Institute of Technology and
Science, Hyderabad, India.
pvssrinivas23@gmail.com

Dr. A Govardhan

Department of CSE, Jawaharlal
Nehru Technological University
Hyderabad, India
govardhan_cse@yahoo.co.in

Abstract— The high-performance computing (HPC) systems participate an significant responsibility in many highly computational applications and systems. Understanding the failure behavior of such a massively parallel system is essential to accomplishing high utilization of large systems. This process requires continuous on-line monitoring and analysis of all incidents generated in the system, including long-term normal notification, performance metrics, and failures. This article illustrates the significance of HPC-Ss (HPC-S) and their fault tolerance, especially rules-based systems. To explore the efficient Fault-Tolerant (FT) mechanism and fault prediction method for an efficient FTmechanism in distributed systems with different rules. We also analyzed the progress of HPC in the rule-based distributed system and its future development direction.

Keywords- *HPC, Rule-Based, Fault tolerance, Distributed Systems*

I. INTRODUCTION

Due to massively parallelization, the complication of computer architectures has made the process of developing High-performance computing (HPC) systems very challenging. The HPC systems (HPC-S) participate an important responsibility in today's society. Widespread applications consist of "weather forecasting", "aircraft collision simulation", "computational fluid dynamics in aerodynamic studies", "bioinformatics", "molecular modeling of protein folding in biomedical research", etc. [2], [3], [4], [5]. However, many of the procedural confronts that HPC-S face as they produce to higher levels because of the amplified numeral of processors complicated, due to the indicate instance among failures of individuality structure sections and the represent time between malfunction of the entire system [7], [9], [10]. As programs and information spread transversely "hundreds of thousands of different processors" and "separate memory libraries", the organization of data exchange and sequential computation necessitates extremely composite reason, a large quantity of dedicated training and fault tolerance. The majority programs depends on a messaging documentation that necessitates moderately short logic and commands. It may require concentrated debugging and optimization tools production.

Cloud Computing delivers innovative computational concepts, capability, and agility to "HPC applications" by

providing numerous "virtual machines (VMs)" for compute-demanding applications that use cloud services. It is likely that compute-demanding applications determination be installed more and more in HPC-S in cloud computing [15], [16]. For instance, an "Amazon Elastic Compute Cloud (Amazon EC2)" [17] cluster.

The key challenges facing HPC-S mentioned in the literature can be separated into (1) fault tolerance and (2) rollback recovery. These are two major issues facing the HPC-S. The reason of this learning is to improve the fault tolerance of HPC-S through rule-based prediction and a new rollback recovery scheme. The current research is studied in detail.

A. Fault Tolerance

Current research by Schroders and Gibson [8], Eggetoh et al. [11] and Yigitbasi, et al. [12] and the data collections make available in [25] demonstrate that HW such as "processor, hard drive, integrated circuit socket, and memory", caused additional 50% of malfunction on HPC-S. These works moreover demonstrate:

- The malfunction rate is approximately comparative to the many CPUs.
- The strength of the efforts affects the failure rate [8].
- There is a association with the malfunction rate over time [12], [13].

B. Rollback-Recovery

The "Rollback Recovery Fault Tolerance (RR-FT) techniques" are frequently utilized by the HPC-S community [14], [19]. When individual or more computational nodes fail, they tend to reduce the impact of the malfunction on "compute-intensive applications" that execute on HPC-S. A high-quality illustration of RR-FT is "checkpoint" and "restart". Checkpoints and restarting allow computationally demanding difficulties to facilitate to acquire extended period to accomplish on HPC-S to restart commencing a checkpoint in the occurrence of an error or failure.

Nevertheless, current publications [1], [2], [4], [10], [16] show that as the numerous elements in nowadays HPC-S are prolongs to develop and the applications operation on HPC-S possibly will not be capable to communicate with fundamental

checkpoints and restart the method to progress. It is for the reason that the system will utilize mainly of the instance on the checkpoint, which is not element of the calculation activity. It is especially important that FT solutions will diminish the overhead of rollback recovery [35].

The following article reviews the significance of HPC in the second part. The FT mechanism in HPC-S is discussed in Section 3, and the rule-based HPC distributed system is discussed in Section 4. In the 5 and 6 sections, the related work was investigated and the conclusion of the evaluation was summarized.

C. The exploit of RTT to Infer Congestion

However, the Internet needs to provide some outline of feedback to data traffic originating from the congested links so that it can regulate its transmission rate depending on the accessible bandwidth, effectively managing end-to-end congestion control [16]. Feedback on congestion can be implicit or explicit. In the case of implicit response, the network's transport layer protocol attempts to maintain high throughput by approximation "service time", "end-to-end delay", and "packet deliver fail". The TCP protocol widely used by the Internet [7], [10] implicitly feeds back lost packets over time and repeatedly. Terminal nodes usually deploy explicit feedback. However, relying on end nodes for implicit or explicit feedback is not enough to achieve a high throughput of the Internet.

II. SIGNIFICANCE OF HPC

The HPC-S dates in past to the 1960s whilst it utilized "parallel and distributed computing" to accomplish high computational concert. The "Parallel computing" utilizes contribute to memory to substitute an information among processors, whereas "distributed computing" utilizes distributed memory to share information among the processors through messaging. In recent times, as parallel computers tend to have some distributed features, it is difficult to distinguish among the parallel and distributed systems.

Currently, "parallel and distributed computing systems" through a huge amount of processors be often referred to as HPC-S. The HPC-S extent from hundreds of processors to hundreds of thousands of processors, for instance the "Clay Titan" [22]. For a moments referred to as "long-executing applications" and "compute-intensive applications" be frequently "scientific calculations" that analyze and solve large and complex scientific problems using "mathematical models" and "quantitative analysis techniques" executing on HPC-S. With cloud computing, HPC-S are no longer confined to huge organizations but moreover to characters. The "Cloud computing" [15] has many advantages, together with no up-front speculation in the procure and setting up of tools and software. The HPC-S in the cloud is a high-quality substitute for conventional HPC-S.

The FT is a system usually computer-based property that continues to operate normally in the occurrence of some component failures [23]. The HPC-S really wants FT because it ensures that "compute-intensive applications" are concluded in

time. In various FT systems, a mixture of one or additional techniques is utilized.

III. FAULT TOLERANCE MECHANISMS IN HPC SYSTEMS

The HPC-S FT is main key confronts facing HPC application in cloud services. There is evidence that a system with "100,000 processors" experiences a processor breakdown each hours [18]. A breakdown take places whilst a HW element fails and requires to be replaced when a "software component fails", "node/processor is stop", or "force a restart", or the software can not absolute a execution. In this matter, the application that uses the unsuccessful component will not succeed.

In accumulation, HPC applications implemented in the cloud execute on "virtual machines", which are further probable not succeed because of "resource sharing" and "contention" [2], [8], [12], [24], [27]. As a result, "FT technology" is especially significant for "HPC applications" that operate in a cloud environment since FT indicates that operating costs and resource consumption are reduced by eliminating the need to restart long-executing applications from scratch in the occurrence of a malfunction.

The availability of interconnected networks in HPC-S is the foundation for the continued execution of large-scale applications. In the incident of a failure, the interconnect recovery mechanism coordinates complex operations to restore the network connection between nodes. As the size and design complexity of HPC-S increase, the system is susceptible to failure during the performance of the interconnect recovery process. HPC components continue to grow to achieve higher performance and meet the requirement of science and other application users. To reduce the average repair time and increase the availability of these systems, a FT solution is required.

The significance of HPC fault-tolerant systems has been extensively recognized by a variety of research institutions in HPC-S. Various schemes [1], [3], [8], [9], [19] have been proposed to offer FT in HPC-S. In this approach [1], [2], [14] explore redundancy and rollback recovery techniques, respectively. Rollback Recovery [1], [14] is one of the most extensively utilized FT mechanisms in HPC-S. Rollback recovery includes checkpoints, fault detection and recovery/restart [6], [8]. Nevertheless, "rollback recovery" typically enlarges the completing period of HPC applications, growing the resource usage and costs of executing HPC applications in legacy HPC-S and in current distributed HPC-S such as the cloud.

A. Software Failure

It is significant to point out that in most systems, 20-30% of the failures are root cause uncertain. Since the percentage of HW breakdowns is greater than the proportion of undecided breakdowns in all systems and the proportion of software malfunctions is nearly the proportion of undecided breakdowns, we be able to still finish that HW & SW are one of the biggest suppliers to failure. Nevertheless, it cannot conclude that several other source of crash besides "human", "environment", "network" is negligible.

In all-purpose, HW issues are intimately observed and properly supervised through the administration system. In adding up, the HW is simple to analyse than software. The distribution of the number of nodes involved in the failure caused by different types of root causes is also different. Faults with HW root cause propagate to a minimum of node aggregates outside the bounds of a small proportion. In contrast, software failures propagate in large proportions. And, even for the same HW problems, different times can be fixed depending on when it happens. For example, "CPUs", "memory", and "node interconnect" difficulties represented by the variation coefficient of LANL systems have a variability of repair times of 36, 87, and 154, respectively. This shows that convenient are erstwhile features that can lead to a high degree of variability in HW breakdowns. Software breakdowns have comparable behavior.

Likewise, El-Sayed and Schroeder [28] considered the field malfunction data for high-performance concrete systems available in [29]. The failure data study was collected for nine years. They observed a considerable association among the network, environment, and software breakdowns. They moreover examined that convenient was a significant increase in the likelihood of software breakdown after a influence problem happened.

Nagappan et al. [30] of "North Carolina State University" studied on-demand practical computing lab failure log files. The "North Carolina State University's virtual computing lab" runs as a confidential cloud through additional than "2000 computers". In this learning, they examined to facilitate system software played a comparatively negligible function in system failures. According to their conclusion, the majority of documented breakdowns were reasoned through "workload", "license depletion", and "HW failures".

B. Hardware Failure

The "USENIX Association" [24] also released a large number of breakdown data, "Computer Failure Database (CFDR)". It includes failure statistics for "22 HPC-S", together with a entirety of "4,750 nodes" and "24,101 processors" accumulated by "Los Alamos National Laboratory (LANL)" over a nine-year period. Workloads include "3D scientific simulations" of great-scale, long-executing operations that acquire months to comprehensive.

To additional check the breakdown rate, we particular seven HPC-S commencing the breakdown database [25]. The system of choice includes the largest total numeral of CPUs and/or clusters of five compute nodes, a "symmetric multiprocessing (SMP) system" through the utmost number of CPUs, and the only "non-uniform memory access (NUMA)" system in the data deposits.

Schroeder and Gibson [8] explored fault data accumulated at two huge HPC sites: "datasets from LANL RAS" [25], and over a year on a huge "supercomputing system" by means of "20 nodes" and over "10,000 processors" data accumulated throughout the period. Their analysis shows that:

- The average repair time for all faults (regardless of fault type) is regarding 6 hours.

- There is a association among the breakdown rate of the system and the functions it executing.
- Some systems may fail three times in 24 hours.
- The breakdown rate is approximately relative to the amount of processors in the system.

C. Human Error Failure

Oppenheimer and Patterson [33] reported that operative error was one of the single leading reasons of breakdown. According to their description, "Architecture and Dependability of Large-Scale Internet Services", the operational staff have replaced systems such as HW replacement, system reconfiguration, deployment, patches, software upgrades, system maintenance and other failures. They attribute "14-30% of failures" to human inaccuracy.

It follows that approximately every breakdowns in "compute-intensive applications" are caused by "HW failures", "software failures", and "human error". Nevertheless, the only foremost reason why failure is clear is difficult, as the analysis reported above is ongoing:

- Utilize different systems executing different applications;
- Beneath diverse background circumstances, and
- Use dissimilar data correlation phases and process.
- Therefore, effectual fault-tolerant HPC-S have to deal with the HW & SW malfunctions as healthy as human mistakes.

IV. RULE-BASED DISTRIBUTED SYSTEMS FOR HPC

The rule-based expert system's early relative success with a more efficient rule-based reasoning engine has driven the application of rule technology to "distributed computing" and "multi-agent systems". This investigate trend follows the parallel promotion of "rule-based reasoning", as fine as the "distributed computing model". There are at slightest of two inclinations that able to be experiential at this time: (i) improved "inference algorithms" for rules systems by means of "parallel and distributed system" technologies; (ii) the additional descriptive environment of "rule-based languages" than procedural languages towards develop a more complex system of autonomous components called "software agents".

A. Parallel Computing for Rule-based Systems

The evolution of computational representations for "rule-based production systems" is principally correlated to the development of the "RETE algorithm" [6] and its extension to the well-organized identical of regular outlines and operational memory constituents [4], except to the simultaneous processing of rules and the establishment of creation systems From the past half of the 1980s, a vivid research route started in the 1990s. The main result of these studies is the powerful implementation of rules and systems for the development of technology.

The researchers of [16] suggested a innovative parallel architecture for nested parallelism of "forward-link inference algorithms" that take advantage of "rule-based production

"systems" on multi-processor systems. The foremost result of this effort was a marked enhancement in the speediness with which "rule-based production systems" were implemented, articulated as rule cycles / second, and changes in functioning memory of elements per second. Their process deal with the entire stages of the familiar link reasoning cycle through: "matching", "parsing", and "right-hand rule evaluation".

The researchers of [20] proposed an thorough investigation of concurrency calculation methods to advance the development of one or more regulation schemes. The author first considers several of the development of "rule-based systems" and later discuss the following as, (i) "parallel production systems, algorithms" that include parallel rules loops, (ii) "distributed production systems" beneath "distributed control", and "multi-agent production systems" and their associated managing concerns.

The researchers of [21] recommend a "parallel and distributed" versions of the "RETE algorithm" using "master-slave paradigms". The outline corresponding system breaks down into master and slave components to work in comparable. Every element contains a duplicate of the "RETE network". The convention are stimulated in similar by the main modules. As soon as a regulation is formulate active, it transmits the entire activated evidence to an existing needy element for commencement and arrives the result. Consequently, the regulations able to be triggered in parallel with the active calculation distributed between the dependent modules.

B. Agent Reasoning Models for Rule-Based Systems

Earlier work suggested the utilize of "rules-based systems" as the essential interpretation model for agents acting as fraction of a "multi-agent system". With this schemes, every agent in the system consist of a rules mechanism, so its activities can decreased to enforcing "rules-based reasoning". The agent synchronization be able to accomplished by sharing a "functioning memory" or "asynchronous messaging".

In [25] a "multi-agent system" known as "MAGSY" is described, where every agent is a "rules-based system". Agents can correspond asynchronously and serve other agents. The "MAGSY" is actually a generic multi-agent framework. Each "MAGSY agent" is a harmony of evidence, regulations, and services. Agents capable of receive information from other agents to trigger an keep posted of their facts. Agents able to call a services offered through last agents. As a consequence, service implementation able to transform the agent's evidence and regulations. Every "MAGSY agent" achieves "rule-based reasoning" with a "forward chain rule interpreter" depend on the familiar "RETE algorithm".

Researchers [26] and [27] believe "multi-agent production systems", which differ theoretically since "parallel and distributed production systems". Although concurrent creation systems give emphasis to "parallel rules matching", and "ring and distributed production systems" highlight the active distribution of production amongst the main groups in an association and aim at improving "execution performance", "multi-agent production systems" focus on multiple self-

determining production systems acting together Working memory, valuable for their management.

C. HPC for Rule-based Systems

Recent research developments can be experiential when studying the synergies between HPC and rules-based systems and reasoning. In case of, the senior expression of "rule-based language" determines the calculation difficulty of presumption algorithms, thus preventive the probable of "rule-based systems" in applications that necessitate big-extent inference. The availability of HPC opens up innovative potential for "scalable rule reasoning" in distributed systems. HPC-S consist of "supercomputers", "computer clusters", "grids" and, more recently, "cloud computing infrastructures".

It may be the first report [31] that uses the results of parallelization of the deductive database and the accessibility of clusters in parallel to consider the enhancement in "semantic web reasoning". The researchers of [32] recommended a data partitioning method, a "parallel algorithm", and numerous optimizations to the "scalable parallel reasoning" of "materialized OWL repositories". The functioning of this algorithm is supported on "Jena open-source rule reasoning" and experiments on a "16-node computer cluster".

In "MARVIN", a parallel and distributed stage for processing great number of "RDF data", is described in [33] and uses a new policy identified as "split-exchange on loosely coupled peer-to-peer network". The initiative of this scheme is to split the "RDF triples" continually, calculate the closures of every separation in parallel, and after that exchange the separations through switch over among peers. This method proved to ultimately accomplish the integrity of interpretation and proposed an effective scheme called "SpeedDate" for substitute data among the peers.

The "Map-Reduce" is a procedure for writing large-scale data computing responsibilities on big "computer clusters" [34]. The "Hadoop" is an "Apache project" that build ups "open-source software" for reliable, scalable, distributed computing" and provides the "Map-Reduce programming framework". In [36] it demonstrates mechanism to pertain "MapReduce" on "Hadoop" to large-scale in "RDFS" inference.

The "Grids" and "Clouds" are recent appearances of distributed computing, placing a high value on "virtualization" and "software services technologies". The grid is synchronize resource contribution and difficulty-outcomes in a "dynamic, multi-agency virtual organization" [30]. The cloud can be deployed and capitalized with minimal administrative effort, and little is known about the infrastructure that supports it. This part temporarily introduces the function of policy and regulation interpretation in improving grid resources and workflow management. Generally of these outcome also pertain to the cloud computing environment. In [9] author introduced a innovative method for accessible performance in on-demand synthesis grids with pertain "ontology rules". The "Rule-based synthesis" unites multiple original actions into innovative composite actions. In [12] described "WS-CAM", a collaborative perceptual management function in a "rules-based grid" environment.

Finally, the "rule-based approaches" prove to be constructive for employ flexible manage schemes and conclusions that permits grids to use "service level agreement (SLA) management systems" to achieve the quality of service promised by various applications.

D. P2P computing for Rule-based Systems

In a "Peer-to-peer (P2P)" representation of distributed systems in which it "equally weighted" and "directly connected peers" work together through given that sources and services to every former peer. The "P2P systems" have significant purposes in "distributed processing", "distributed content management", and "ubiquitous computing". The grouping of "decentralization of the P2P approach" and the declaration and elasticity of the regulations facilitates the improvement of innovative category of intellectual distributed schemes. Applications are provided in heterogeneous schema mapping and ubiquitous computing.

In [34] author introduces the use of an inference engine to represent and process the meaning of digital library resources in a heterogeneous environment. This approach applies to metadata mapping between heterogeneous schemas in a "P2P-based digital library". The Mappings are characterize by taken out facts from the resource's "XML metadata" and then pertaining the "rule-based reasoning" to mechanically extract relationships among the confined schema and erstwhile discovered representations.

In [22] commences a "Prolog's P2P extension", "LogicPeer". In the "LogicPeer" it is a simple addition of "Prolog" using an operator that can perform target assessment on peers in a "P2P system". In "LogicPeer" it describes two network representations. (i) an "opaque peer network model" where every peer does not recognize the identifier of the neighbor router and the specific "query propagation protocol", (ii) each peer gets its neighbor's identifier, thus allowing the implementation of a customized query propagation protocol.

In [25] and [27] a distributed interpretation resolution is proposed that can utilized as a "P2P network environment". Each agent has a partial environmental perspective and has a locally consistent theory. Local theory is linked through bridging rules, which can lead to inconsistencies in the global knowledge base. Handling inconsistencies are done by creating a bridging rule with unmanageable logic.

V. INVESTIGATION OF THE RELATED WORKS

Various researchers have keen to the significance of failing data analysis and the necessitate for an open failed data store. In this white paper, many failed data collected over the last decade on HPC sites have been studied and made publicly available [1].

J. Villamayor et al. [1] presents the "Fault Tolerance Manager (FTM)" for the coordinated checkpoint file to provide automatic recovery from failure when losing the coordinate node. This proposal makes the "Fault Tolerance (FT)" configuration simpler and more transparent for users without knowledge of application implementation. In addition, system administrators do not need to install libraries in the cluster to support FTM. Leverages node local storage to store

checkpoints and distributes copies of nodes along all compute nodes to prevent centralized storage bottlenecks. This approach is especially useful in IaaS cloud environments where users have to pay for centralized, reliable storage services. This is based on "RADIC", a well-known architecture that provides FT in a distributed, flexible, and automatically scalable manner. Experimental results demonstrate the benefits of the approach presented in "Amazon EC2", a private cloud and well-known cloud computing environment.

Saurabh Jha et al. [2] has featured Cray's largest Gemini network Gemini Internet recovery procedure featured at "Blue Rock, a 13.3 petaflops supercomputer" at the "National Center for Computational Applications (NCSA)". It presents a propagation model that captures the failure and recovery of interconnects to help understand the types of faults and their propagation in systems and applications during recovery. The measurement results show that the recovery process is very frequent and the recovery process is unsuccessful in the incident of an additional failure during recovery, resulting in system-wide interruptions and application failures.

I. A. Sidorov [3] focuses on the development of models, methods, and tools to improve the FT of HPC-S. The models and methods described are based on the use of automated diagnostics, automatic localization, error correction, and automatic HPC-S reconfiguration mechanisms for the underlying software and HW components of these systems. The uniqueness and novelty of the proposed approach allow agents to make decisions directly by creating a multi-agent system using a general-purpose software agent that can collect node state data for analysis.

S. Hukerikar et al. [4] proposes a partial memory protection scheme based on area-based memory management. This defines a local concept called havens, which provides error protection for program objects. It also provides reliability for the area through a software-based parity protection mechanism. U. Lakhina et al. [5] The centralized load-balancing algorithm proposed an algorithm that dynamically balances the load and ensures the overall performance of the system. This concept focuses on building FT systems by achieving high resource utilization, reducing job rejection rates, and making improved calculations and backups. The results of this cluster-oriented load-balancing algorithm show that response time is reduced and processing time is improved.

We encourage this data to be the first action toward a open data store and support former sites to accumulate and organize data for open disclosure. Underneath we review some outcomes.

- Failure rates vary widely from system to system, and there are more than 20 to 1000 occurrences each year, depending on system dimension and HW type.
- Since the error rate is approximately relative to the quantity of processors in the system, the error rate does not increase much faster in proportion to the system dimension.
- There is confirmation that there is a relationship among the breakdown rate of the system and the kind and strength of

- the executing jobs load. This is consistent among the initial job on another type of system.
- The curve of the breakdown rate above the life span of the HPC-S is frequently extremely dissimilar from the life cycle curve details in the background study of individual HW or SW modules.
- The occasion among failures is not well created through an "exponential distribution" that matches the initial discovery for other types of systems. Identify the time to failure of the entire system as well as individual nodes.
- The average repair instance fluctuate from "one to more than one hour", depending on the system. The repair time is largely dependent on the type of system and is not affected by the size of the system.
- Restore occasions are highly inconsistent inside a system and are a large amount of enhanced formed through logarithmic normal rather than the "exponential distribution".

This survey task has difficulties in solving key challenges and problems that can be improved to provide new solutions in the field of HPC-S. The key enhancements should focus on the FT framework of HPC [3], [9] that able to utilized to construct consistent HPC-S using rule-based error prediction methods. Investigate the impact of various parameters on overall forecast results and highlight the best combination of the log file and the described incidents of failure. Even the implementation of a hybrid FT approach that combines fault prediction and multi-step checkpoint techniques can overcome system overhead for many HPC applications [1], [6], [8], [19].

VI. CONCLUSION

High-performance computing (HPC) systems can use interconnection networks, allowing applications to execute continuously, regardless of size. In the incident of a failure, the interconnect recovery mechanism coordinates complex operations to recover network connectivity between nodes. As the size and design complexity of HPC-S increases, the system may fail during the interconnect recovery procedure. This review focuses on insights into HPC components and continues to grow to meet performance and user needs for science and other applications. It focuses on the importance and FT mechanisms and causes of HPC. Later, we discuss related works related to the function of HPC in "rule-based distribution systems". Understanding the challenges of HPC and the research support that can be improved in the future are needed to reduce mean time to recovery and propose new fault tolerance solutions in these systems to increase high availability.

REFERENCES

- [1] J. Villamayor, D. Rexachs, E. Luque, "A Fault Tolerance Manager with Distributed Coordinated Checkpoints for Automatic Recovery", International Conference on High Performance Computing & Simulation (HPCS), pp. 452 - 459, 2017.
- [2] S. Jha, V. Formicola, C. Di Martino, M. Dalton, et al., "Resiliency of HPC Interconnects: A Case Study of Interconnect Failures and Recovery in Blue Waters", IEEE Transactions on Dependable and Secure Computing, Volume: PP, Issue: 99 pp. 1 - 1, 2017.
- [3] I. A. Sidorov, "Methods and tools to increase fault tolerance of high-performance computing systems", 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 226 - 230, 2016.
- [4] S. Hukerikar, C. Engelmann, "Havens: Explicit reliable memory regions for HPC applications", IEEE High Performance Extreme Computing Conference (HPEC), pp. 1 - 6, 2016
- [5] U. Lakhina, N. Singh, A. Jangra, "An efficient load balancing algorithm for cloud computing using dynamic cluster mechanism" 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1799 - 1804, 2016.
- [6] I. Gankevich, Y. Tipikin, V. Korkhov, V. Gaiduchok, "Factory: Non-stop batch jobs without checkpointing", International Conference on High Performance Computing & Simulation (HPCS), pp. 979 - 984, 2016.
- [7] D. W. Kim, M. Erez, "Balancing reliability, cost, and performance tradeoffs with FreeFault", IEEE 21st International Symposium on High Performance Computer Architecture (HPCA) pp. 439 - 450, 2015.
- [8] B. Schroeder and G. Gibson, "A large-scale study of failures in high-performance computing systems", IEEE Transactions on Dependable and Secure Computing, vol. 7, no. 4, pp. 337-350, 2010.
- [9] I. P. Egwuatuoha, S. Chen, D. Levy, and B. Selic, "A Fault Tolerance Framework for High Performance Computing in Cloud", in 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 709-710, IEEE Press, 2012.
- [10] R. Riesen, K. Ferreira, and J. Stearley, "See applications run and throughput jump: The case for redundant computing in HPC", in 2010 International Conference on Dependable Systems and Networks Workshops (DSN-W), pp. 29-34, IEEE Press, 2010.
- [11] I. P. Egwuatuoha, D. Levy, B. Selic, and S. Chen, "A survey of fault tolerance mechanisms and checkpoint/restart implementations for high performance computing systems", in The Journal of Supercomputing, vol. 65, pp. 1302-1326, Springer, 2013.
- [12] N. Yigitbasi, M. Gallet, D. Kondo, A. Iosup, and D. Epema, "Analysis and modeling of time-correlated failures in large-scale distributed systems", in 2010 11th IEEE/ACM International Conference on Grid Computing (GRID), pp. 65-72, IEEE Press, 2010.
- [13] N. Xiong, M. Cao, J. He and L. Shu, "A Survey on Fault Tolerance in Distributed Network Systems", Proceedings of the 2009 International Conference on Computational Science and Engineering, Vancouver, 29-31, 1065-1070, 2009.
- [14] E. N. Elnozahy, L. Alvisi, Y.-M. Wang, and D. B. Johnson, "A survey of rollback-recovery protocols in message-passing systems", ACM Computing Surveys (CSUR), vol. 34, no. 3, pp. 375-408, 2002.
- [15] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, et al., "A view of cloud computing", Communications of the ACM, vol. 53, no. 4, pp. 50-58, 2010.
- [16] C. Evangelinos and C. Hill, "Cloud Computing for parallel Scientific HPC Applications: Feasibility of running Coupled Atmosphere-Ocean Climate Models on Amazon EC2", ratio, vol. 2, no. 2.40, pp. 2-34, 2008.
- [17] Amazon EC2, "Amazon Elastic Compute Cloud (Amazon EC2)." <http://aws.amazon.com/ec2/>. Online: accessed in April, 2013.
- [18] A. Geist and C. Engelmann, "Development of naturally fault tolerant algorithms for computing on 100,000 processors", Submitted to Journal of Parallel and Distributed Computing, 2002.
- [19] J. C. Sancho, F. Petrini, K. Davis, R. Gioiosa, and S. Jiang, "Current practice and a direction forward in checkpoint/restart implementations for fault tolerance", in 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05), p. 300.2, IEEE Press, 2005.

- [20] F. Cappello, A. Geist, B. Gropp, L. Kale, B. Kramer, and M. Snir, "Toward exascale resilience", International Journal of High Performance Computing Applications, vol. 23, no. 4, pp. 374-388, 2009.
- [21] F. Cappello, "Fault tolerance in petascale/exascale systems: Current knowledge, challenges and research opportunities", International Journal of High Performance Computing Applications, vol. 23, no. 3, pp. 212-226, 2009.
- [22] H. W. Meuer, and others., "TOP500 Supercomputer Sites." <http://www.top500.org/>. Online: accessed in April, 2013.
- [23] Wikipedia.org, "Fault-tolerant system." http://en.wikipedia.org/wiki/Fault_tolerant_system. Online: accessed in April, 2013.
- [24] USENIX, "Computer failure data repository (cfdr)." <https://www.usenix.org/cfdr>. Online: accessed in April, 2013.
- [25] Arif Sari, Murat Akkaya, "Fault Tolerance Mechanisms in Distributed Systems", Int. J. Communications, Network and System Sciences, Vol. 8, PP. 471-482, 2015
- [26] C. D. Martino, S. Jha, W. Kramer, Z. Kalbarczyk, and R. K. Iyer, "Logdiver: a tool for measuring resilience of extreme-scale systems and applications", in Proceedings of the 5th Workshop on Fault Tolerance for HPC at eXtreme Scale, pp. 11-18, ACM, 2015.
- [27] J. B. Leners, H. Wu, W.-L. Hung, M. K. Aguilera, and M. Walish, "Detecting failures in distributed systems with the falcon spy network", in Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, pp. 279-294, ACM, 2011.
- [28] N. El-Sayed and B. Schroeder, "Reading between the lines of failure logs: Understanding how HPC systems fail", in Dependable Systems and Networks (DSN), 2013 43rd Annual IEEE/IFIP International Conference on, pp. 1-12, IEEE, 2013.
- [29] Los Alamos National Laboratory., "Operational Data to Support and Enable Computer Science Research". <http://institute.lanl.gov/data/fdata/>, 2014. Online: accessed in June, 2014.
- [30] M. Nagappan, A. Peeler, and M. Vouk, "Modeling cloud failure data: a case study of the virtual computing lab", in Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing, pp. 8-14, ACM, 2011.
- [31] C. Di Martino, F. Baccanico, J. Fullop, W. Kramer, Z. Kalbarczyk, and R. Iyer, "Lessons learned from the analysis of system failures at petascale: The case of blue waters", in Proc. of 44th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN), 2014.
- [32] V. Puente, J. A. Gregorio, F. Vallejo, and R. Beivide, "Immunet: A cheap and robust fault-tolerant packet routing mechanism", in Computer Architecture, 2004. Proceedings. 31st Annual International Symposium on, pp. 198-209, IEEE, 2004.
- [33] D. Oppenheimer and D. A. Patterson, "Architecture and dependability of large-scale internet services", IEEE Internet Computing, vol. 6, no. 5, pp. 41-49, 2002.
- [34] M. Balazinska, H. Balakrishnan, S. Madden and M. Stonebraker, "Fault-Tolerance in the Borealis Distributed Stream Processing System. ACM Transactions on Database Systems, 33, 1-44, 2008.
- [35] E.N.M. Elnozahy, L. Alvisi, Y.M. Wang and D.B. Johnson, "A Survey of Rollback-Recovery Protocols in Message-Passing Systems. ACM Computing Surveys, 34, 375-408, 2002.
- [36] D. Tian, K. Wu and X. Li, "A Novel Adaptive Failure Detector for Distributed Systems", Proceedings of the International Conference on Networking, Architecture, and Storage, Chongqing, PP. 215-221, 2008.

Risk Assessment: Approach to enhance Network Security

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem, Okonkwo, Obikwelu Raphael

¹Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
Nwaguchikeziekenneth@hotmail.com

²Computer Science Department, Michael Okpara University of Agriculture Umudike Umuahia, Abia State, Nigeria.
Saintbeloved@yahoo.com

³Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
Oobi2971@yahoo.com

Abstract

This research work x-rays the indispensability of continuous risk assessment on data and communication devices, to ensure that full business uptime is assured and to minimize, if not completely eradicate downtime caused by “unwanted elements of the society” ranging from hackers, invaders, network attackers to cyber terrorists. Considering high-cost of downtime and its huge business negative impact, it becomes extremely necessary and critical to proactively monitor, protect and defend your business and organization by ensuring prompt and regular Risk assessment of the data and communication devices which forms the digital walls of the organization. The work also briefly highlights the methodologies used, methodically discusses core risk assessment processes, common existing network architecture and its main vulnerabilities, proposed network architecture and its risk assessment integration(Proof), highlights the strengths of the proposed architecture in the face of present day business threats and opportunities and finally emphasizes importance of consistent communication and consultation of stakeholders and Original Equipment Manufacturers (OEMs)

Keywords- *Risks,Risk Assessment,Network Architecture,vulnerabilities,Opportunities.*

I. METHODOLOGY

In the course of this work, some other methodologies such as Object Oriented Analysis and Design Methodologies (OOADM), Incremental/Evolutionary Methodologies etc. were considered and finally adopted Structured System Analysis and Design Method(SSADM), Dynamic System Development Method (DSDM) and Spiral Methodologies.

Reasons for the adoption are their direct applicability and features among which are respectively as follow:

For SSADM,

- Intensive users involvement
- Clear and easily understandable documentation
- Process is Procedural

For DSDM

- Focuses on Business need and delivers on time

- Communicate continuously and clearly without compromising quality
- For Spiral

- Risk driven and keeps track of risk patterns in a project.
- Iterative and incremental

All these features are used to analyze common existing network architectures with the primary aim of:

- Identifying bottlenecks and Problems thereof.
- Investigating areas of improvement
- And proffering solutions to the system(Proposed Network Architecture)

II RISKS AND RISK ASSESSMENT

If you don't assess your data and communication devices, definitely someone else would. And this will invariably leave your organization at the mercy of the attackers – attackers are ill-winds that blow no man any good. In fact, in most cases, organizations are shutdown, monies are lost and the image of the company is battered and left in doubt for the public as data integrity, confidentiality and availability are not assured of.

Risks, contrary to earlier notion, are both positive and negative. Therefore Risk, as adapted from Stoneburner, G., Goguen, A. & Feringa, A. (2002, July), is net negative or positive effect of exercise of vulnerabilities or opportunities which can be exploited, enhanced, shared, transferred, or even accepted.

However, this work focuses on negative risks (vulnerabilities) which can be intentionally exploited or accidentally triggered. Subsequent publication which is part of the entire work, will atomize risks as opportunities which are positive.

Risk Assessment, as a critical part of risk management, is made of many core processes (which the steps depicted in figure 3 and further explained) such as:

- Risk identification: This allows individuals to identify risks so that the stakeholders will be in the know of potential threats or opportunities inherent in the devices. It is pertinent to start this stage as early as possible and should be repeated frequently.

- Risk analysis and Prioritization: Risk analysis transforms the estimates or data about specific risks that developed during risk identification into a consistent form that can be used to make decisions around prioritization. Risk prioritization enables operations to commit resources to manage the most important or worst risks.
- Integration of Risk registers: This assures that risks (both low and high priority ones) are tracked and monitored through the entire process. In the course of the initial process of Risk assessment, all low-priority risks are kept in the register and while during subsequent risk assessments, the content is updated after the content is assessed, in addition to the entire assessment. As a core part of process and result, contains lists of identified risks, root causes of risks, lists of potential responses, risk owners, symptoms and warning signs, relative rating or priority list. These are risks for additional analysis and responses, and a watch list which is a list of low-priority risks under close watch and monitoring.
- Communication and Consultation: Communication is key in risk assessment. There should be steady and consistent communication/consultation among stakeholders within the organization as everyone is practically involved and outside, especially to Original Equipment Manufacturers (OEMs). In addition, the stakeholders can utilize all their communication channels (Calls, Emails, via Technical Account Managers (TAM), Portal etc.) to the OEMs. This ensures speedy and reliable responses which further assist organizations to strategically align with industry best practices and proactively avoid negative risks.

III. RISK ASSESSMENT KEY STEPS

The Principal steps undertaken in order to thoroughly assess risks associated with ICT systems are:

- Step 1 – System/Device Characterization
- Step 2 - Vulnerability Identification
- Step 3 - Threat Identification
- Step 4 - Control Analysis
- Step 5 - Likelihood Determination
- Step 6 - Impact Analysis and Integration of Risk register
- Step 7 - Risk Determination
- Step 8 - Control Recommendations
- Step 9 - Assessment Documentation and Updating Risk Register

Depending on the level of organizations risk awareness and maturity, steps 2 to 6 can be carried out concurrently after completing step 1. And throughout the entire processes and steps ,there is consistent two-way communication and consultation among stakeholders including OEMs. This is key to ensure successful and seamless risk assessment processes integration into organizational daily operational routines

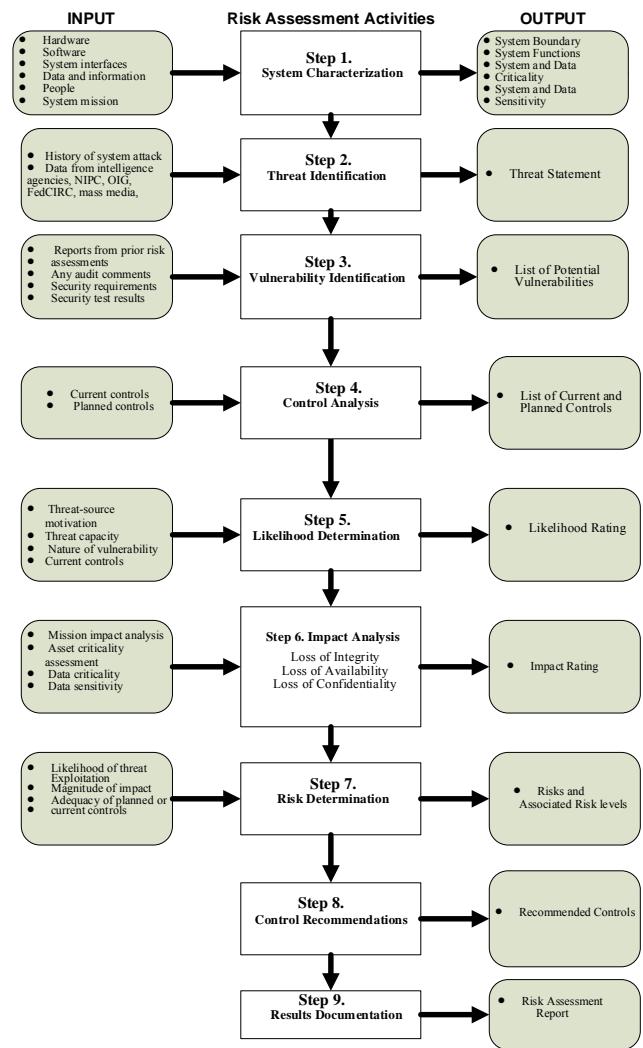


Figure 3: Risk Assessment Process (*adapted from stoneburner et al.*)

Step 1 – Device/System Characterization:

The first step, is basically to define the scope of ICT systems, be it data and communication devices, hardware, software or as your organization need directs. This would assist the stakeholders to establish clear category as outputs at this step based on functionality, criticality and /or sensitivity of the devices/systems. With these, it will be essentially easy to define the risk, its priority and of course, assign appropriate resource which would help to mitigate the identified risks completely at this stage or reduce it to acceptable level which would immediately be kept in the risk register for close monitoring and watch in the subsequent risk assessments. If this happens, would automatically jump the process to step 7 which keeps the end in sight for the entire processes.

Step 2 - Vulnerability Identification

This is principally looking at the existing documentations (risk assessment etc.) and records (audits, continuous quality/process improvement etc.) and studying them, so as to

come up with list of vulnerabilities and /or potential vulnerabilities associated with respective devices/ICT systems.

Step 3 - Threat Identification

This involves studying history of attacks and possibly their trend. Accessing and leveraging on similar information from other organizations would be an added advantage as all known threats and their respective sources would reflect on the threat statement. With this broader coverage of threat statement, the risk assessment core processes would be on the lookout for them and mitigating plan proactively put place to ensure no exercise of those threats by the threat sources.

Step 4 - Control Analysis

Having established the systems boundary, possible vulnerabilities identified and threat statement clearly written, it becomes important to analyze existing controls and their respective efficacies and potencies against them(identified vulnerabilities and threat-sources).With these, list of existing current controls would be written and their inadequacies bridged as planned(proposed) controls.

Step 5 - Likelihood Determination

At this stage, threat-source motivation should have been known and put into consideration, nature of the vulnerability and of course, threat capacity if successfully exercised, then likelihood rating could be determined. Likelihood rating is simply the probability that identified vulnerabilities can be exercised successfully by threat-sources. To determine the likelihood involves examining threat-source motivation and capability, type of vulnerability involved and effectiveness of existing controls. Hence the likelihood that a particular potential system weakness could be exercised by a given threat-source is then categorize as **high or low**. However in most cases, organizations go a little deeper and categorize it as **high, medium, or low** depending on the anticipated severity.

It is High when the threats are likely exploitable and the threat source is attracted, very motivated and highly capable to initiate it. However the existing controls are ineffective and insufficient to prevent exercise of the vulnerability thereof.

It is Medium when threats are highly exploitable, the threat source attracted, very motivated and highly capable. However the existing controls may prevent exercise of the vulnerability.

It is low when the threats are not likely exploitable, threat source is not attracted and lacks motivation and capability to initiate the attack. This is mainly due to the existing controls which are sufficient to prevent the exercise of the vulnerability.

Step 6 - Impact Analysis and Integration of Risk register.

This is one of the critical steps of risk assessment processes as the probable adverse impact of successful exercise of the vulnerable is determined and measured for negative risks and optimal benefits/full utilization of the ICT systems is also determined for positive risks to justify for stakeholders especially senior management team the decision taken as regards to likely return on investment (ROI). Integration of

Risk register at this point in the process is another remarkable strength as it assist to build comprehensive list of low impact risks as the entire impact analysis is being conducted.

According to Kosutic, (2014) the purpose of this analysis is primarily to give one an idea:

- a) About the timing of your recovery, and
- (b) The timing of your backup, since the timing is crucial – the difference of only a couple of hours could mean life or death for certain organizations if hit by a major incident. For example, if it is a financial institution, recovery time of four hours could mean you will probably survive a disruption, whereas recovery time of 12 hours is unacceptable for certain systems/activities in a bank, and disruption of a full day would probably mean such a bank would never be able to open its doors again. And there is no magic standard which would give you the timing for your organization – not only because the timing for every industry is different, but also because the timing for each of your activities could be different. Therefore, you need to perform the (business) impact analysis to make correct conclusions for likely successful exercise of vulnerabilities.

Hence Business Impact analysis is to evaluate the impact of the affected ICT systems/devices to the business and entire organization which could be loss of integrity, loss of confidentiality or loss of availability.

According to Chia,(2012), CIA refers to Confidentiality, Integrity and Availability as Confidentiality of information, integrity of information and availability of information. Many security measures are designed to protect one or more facets of the CIA triad.

Exercise of vulnerabilities result could result in entire bleach of security goals (**integrity, availability and confidentiality**) or any of their combinations. Stoneburner et al, explained the security goals in terms of system and data as follows:

- a) **Loss of Integrity.** System and data integrity refers to the requirement that information be protected from authorized modification. Integrity is lost if unauthorized changes are made to the data or IT system by either intentional or accidental acts. If the loss of system or data integrity is not corrected, continued use of the contaminated system or corrupted data could result in inaccuracy, fraud, or misleading decisions. Also, violation of integrity may be the first step in a successful attack against system availability or confidentiality. For all these reasons, loss of integrity reduces the assurance of an IT system.
- b) **Loss of Availability.** If a mission-critical IT system is unavailable to its end users, the organization's mission may be affected. Loss of system functionality and operational effectiveness, for example, may result in loss of productive time, thus impeding the end users' performance of their functions in supporting the organization's mission.
- c) **Loss of Confidentiality.** System and data confidentiality refers to the protection of information

from unauthorized disclosure. The impact of unauthorized disclosure of confidential information can range from the jeopardizing of national security to the disclosure of Privacy Act data. Unauthorized, unanticipated, or unintentional disclosure could result in loss of public confidence, embarrassment, or legal action against the organization.

In view of these, magnitude of impact of successful exercise of vulnerabilities could be classified as just ***low or high; high, medium or low*** depending on the risks threshold or appetite of organizations. The classification could also be done along the magnitude of anticipated loss. Finally, this step comes out with impact ratings for probable successful exercise of identified vulnerabilities. Of course, the risk register is consequently updated accordingly and stakeholders are timely communicated of the impact, its ratings and risks under close monitoring (content of the risk register)

Step 7 - Risk Determination

At this step, for the risk to be accurately determined, steps 1 to 6 are put into consideration specifically its respective inputs and outputs. With these, likelihood of vulnerability exploitation, magnitude of the impacts and effectiveness and adequacy of planned and existing controls are thoroughly examined. As output, all the identified risks and associated risk scale (***low, medium or high***). Significant result of this step is the risk matrix which usually 3×3 (though some organizations may decide to go more granular by adopting 4×4 or 5×5). The matrix is a logical consideration of likelihood of exploitation of the identified vulnerabilities by threat sources and the associated impact if successful which implicitly has considered the existing and planned controls.

Step 8 - Control Recommendations

This is a critical step in the risk assessment process as the consultative and communication channels among stakeholders and the OEMs are further utilized to ensure that commensurate, trendy and best of security controls are recommended to fully mitigate or eliminate existing identified vulnerabilities. With the best of security controls, intending threat sources are totally discouraged or arrested in the course as the costs of successful exploitation of vulnerabilities are made far higher than the anticipated gain thereby discouraging and checking attackers . Hence the output of this step is a list of recommended control in addition to existing controls.

Step 9 - Assessment Documentation and Updating Risk Register

As the final step, all the risk assessment steps and processes are documented including the identified vulnerabilities, impact of successful exploitation, list of existing and recommended controls etc. are presented to the process owners and stakeholders. It is advisable for organizations to create a platform (may be an intranet portal for stakeholders only) on which these reports are timely shared. This is important too, as

risk assessment responsibility is for all stakeholders and not “esoterically” reserve for ICT experts as the practice has always been.

IV. COMMON EXISTING NETWORK ARCHITECTURE AND ITS MAIN VULNERABILITIES

A common typical existing Network architecture is flat and vulnerable as all the data and communication devices such as routers, Idirect, Pixs, switches, fortigate, Pix and many other devices including those from internet Service Providers. These devices are connected directly to the organization’s core switch and/or mostly via the organization’s router. Even the servers, PABX and other telecommunication devices are all connected directly to one and same core switch. Of course, the stacked switches for end users’ devices are linked up and also connected to the same core switch.

This design does not only give room for one point of failure(the core switch) for the entire network but any eventual and slightest break-in through any of these devices comfortably drops the attacker into the heart of the organization’s business data and the rest could be story as your guess of the result is as good as mine.

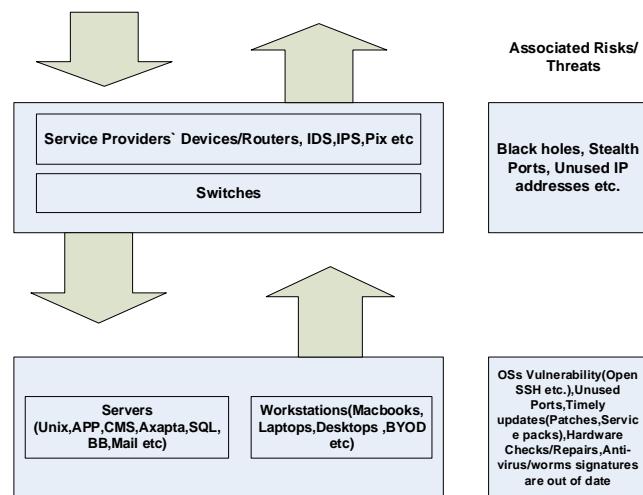


Figure 1: Common Traditional network architecture.

Above block diagram clearly illustrates existing network architecture as implemented by many organizations. The arrows going into the network show traffic from the internet as requested and originated by users in accordance to business needs and the ones going out of the network are traffics sent out by users to fulfill same and/or different business needs.

The first layer shows boundary devices which comprises of the Routers, Idirect, IPS, IDS, Pix, switches and many others as the case may be.

The second layer shows servers(mail,Application,Web server/proxy,DCs,File,Hyper-Vs,DPM,SQL, Axapta,CM etc.), devices such as WAN optimizer,PABX, and end-users’ devices such as the laptops, desktops,BYODs etc., of course, its associated risks/threats at the level.

And most organizations allow remote access such as telnet etc. on the core switch for easy administration of the network from home or out of office, without switch port security, creation of VLANs for the used ports and shutting down unused ports. The risks thereof ranges from Black holes, unused ports, granting access to a range of IP addresses while only few are used, OS vulnerabilities, delayed update of anti-virus and worm signatures etc. This architecture is not a total proof against attacks as a sudden access into any of the boundary devices will totally expose entire network and leaves it at the mercy of the attackers. Again there is no obvious communication and consultation among the stakeholders, not to talk of, with the OEMs

The main Vulnerabilities of the existing network architecture are:

- Being a Flat architecture - all boundary devices are connected to one device and same LAN(no segregation) which makes it easy for an attacker who finds his way into network (router, switch, Idirect, etc.) accidentally or intentionally to fully exercise his attack and bring down the entire network. Also this gives room for one point of total failure for entire network as earlier hinted.
- Granting access to IP address range: If free IP address is sniffed and entrance is successful, routes will be changes and other configurations altered.
- According to Pascucci M.(2012), Router Misconfigurations such as
 - HTTP Open on the core Router: With this, the routes could be changed and NVRAM wiped.
 - Password files stored on Router: Most admins stored company's credentials on a file on the router's storage. Most of the routers run SSHv1 and penetration tests gain access to the file which offers limitless access of the company to the invader. Although admins "believe" that .doc cannot be opened on Cisco IOS.
 - Allow Telnet and other remote accesses open on the boundary devices (routers, switches etc.): Since this is a flat network and telnet is opened to the LAN on the core switch and router, any accidental access to a workstation or even through a compromised user credential via VPN grants an attacker access to the devices which automatically makes it vulnerable.
- According to Manes C. (2014), most common misconfigurations such as:
 - Leaving Shutdown in the remote session options: This is mainly applicable in servers and it's a common practice by systems/Network administrators and their "reason" being that they want to be able to

restart and /or shutdown the device if need be remotely

- Leaving sample applications and code on a production system: Sample applications are meant to guide you on how to do something. It has been discovered that system administrators inadvertently or intentionally leave both the sample application and codes on the production systems. These are valuable tools for attackers.
- Autoconfigured IP addresses in DNS: if a server has two ip.addrs in DNS, it will reply to a query with both of them. If one of those addresses is bogus, a client stands a 50:50 chance of trying that bogus address before it tries the legitimate one. This typically means slow performance and call for assistance.
- Dropping ICMP: This makes it impossible to carry out basic connectivity troubleshooting to the device. However, some administrators drops ICMP which is against the RFCs which states that Hosts must respond to ICMP echo requests. This make it difficult to easily ascertain uptime status of the devices.
- DNS Islanding in Active Directory (AD): This occurs when a DC points to itself for DNS instead of to another. This makes it, in most cases to be out of synchrony with other DCs and if it stays out for too long (about 60 days default) it simply means that it would not "talk" to other DCs which makes all the devices that uses it for name resolution will be rendered incommunicado.

V. PROPOSED NETWORK ARCHITECTURE AND ITS RISK PROOF INTEGRATION

The proposed network system architecture has two-layer security which offers proof against all manner of attacks. Figure 4 is a block architecture of the proposed Network Architecture. The two layers are named private and public. The private communication devices boarder the corporate network and are configured with the organization's Private IP addresses with corresponding subnets such as 10.10.x.x/24. All the corporate communication devices are connected to these devices via stack of Private switches.

However, at the public level, all the connectivity from all service providers' devices are terminated on the public switch on which the public router is directly connected too. The Private router is connected to the public switch. Then the public switch is connected WAN ports of a strong firewall device such as fortigate while its LAN port is connected to the core switch of the LAN. And finally one of the interfaces of the Private router is connected to the same core switch of the LAN. From here (core switch) the signals flow down and up to the stack of private (LAN) switches. However, on the public switch, each connected port has port security access configured and any unused port is shutdown to ensure no vulnerability is

exploitable at all. Port security will work on host port. In order to configure port security we need to set it as host port. It could be done easily by switchport mode access command. You can secure trunk connections with port security.

According to Cisco press (2014), the following configuration below illustrates available commands for port security and port shutdown:

To configure port security:

```
Switch> enable
switch#configure terminal
Enter configuration commands, one per line. End with CNTL/Z
switch(config)#interface fastethernet 0/1
Switch (config-if)#switchport port-security ?
mac-address Secure mac address
    Maximum max secure addresses
    <cr>
Switch (config-if)#switchport port-security mac-address ?
    H.H.H 48 bit mac address
    sticky Configure dynamic secure addresses as sticky
Switch(config-if)#switchport port-security violation ?
    protect security violation protect mode
    restrict security violation restrict mode
    shutdown security violation shutdown mode
Switch(config-if)#switchport port-security host
```

To shutdown a switch port:

```
Switch> enable
switch#configure terminal
Enter configuration commands,one per line. End with CNTL/Z
Switch (config)# interface gigabitethernet1/0/2
You can verify your settings by entering the show interfaces
privileged EXEC command.
```

In addition to port security, Telnet and other remote accesses are disabled on both the public and private boundary devices. And on the private devices mainly the core switch, the following features are configured:

- Private VLAN edge: This offers security and logical boundary - isolation between ports on a switch, hence ensures that different packets such as the VOIP traffic travels directly from its entry point to the aggregation device through a virtual path and cannot be directed to a different port.
- Port-based ACLs for Layer 2 interfaces: This is security policy applied on per-Layer 2 port basis. With this, incoming traffics are matched against the ACLs and good matches are allowed passage .And others are denied passage.
- Multilevel security on console access: This ensures that only authorized users allowed to effect changes on the configurations of the devices. Hence prevent unauthorized users from altering the switch configurations. Of course, authorized users have

varying degree of access which limits what a user can do.

- Granting Specific access on the public layer devices such as to specific host or IP address instead of IP address range or network. This is very critical on this layer. However on the private layer, it can be diluted a little to ensure VLANs talk to one another
- Cisco security VLAN ACLs (VACLS) on all VLANs: This functions as logical boundaries which assures that inhibit unauthorized data flows to be bridged within VLANs.
- Operating Systems Updates: All patches should be tested in a test environment before rolling out to production environment. And whenever in doubt, one of the communication channels (as earlier stated) should be activated and followed-up to ensure timely response.
- Engaging Experts with relevant experience: This is also key to ensure that desired security configurations or intentions are achieved according to organization's needs.
- Port-Level Traffic Controls: This is principally configure on the public layer specifically on the public switch. This offers storm control among many other security features.
 - LAN or network Storm happens when excessive hostile packets are sent continuously on the LAN segment creating unnecessary and excessive traffics which results in network performance degradation. The storm control prevents disruption to regular and normal traffics which is mainly causes by multicast, broadcast or unicast packet storm on any of the physical interfaces.
- To enable traffic storm control feature, at the global configuration mode, use **storm-control {broadcast / multicast / unicast}**. However, by default it is disabled. The storm-control action **{shutdown | trap}** command is used to instruct the switch the action to take when storm is detected. By default, the storm is suppressed which means that no action is configured.
- To verify/check the storm-control suppression levels on an interface, use **show storm-control [interface] [broadcast / multicast / unicast]** command.
- Additional Layer 2 best security practices: This is highly recommended to ensure that best of layer 2

security measures are in place in the private layer switches(core and other stacked switches).According to Bhajji Y.(2008),best practice for managing, implementing and maintaining secure layer 2 network are:

- Manage the switches in a secure manner. For example, use SSH, authentication mechanism, access list, and set privilege levels.
- Restrict management access to the switch so that untrusted networks are not able to exploit management interfaces and protocols such as SNMP.
- Always use a dedicated VLAN ID for all trunk ports.
- Be skeptical; avoid using VLAN 1 for anything.
- Disable DTP on all non-trunking access ports.
- Deploy the Port Security feature to prevent unauthorized access from switching ports.
- Use the Private VLAN feature where applicable to segregate network traffic at Layer 2.
- Use MD5 authentication where applicable.
- Disable CDP where possible.
- Prevent denial-of-service attacks and other exploitation by disabling unused services and protocols.
- Shut down or disable all unused ports on the switch, and put them in a VLAN that is not used for normal operations.
- Use port security mechanisms to provide protection against a MAC flooding attack.
- Use port-level security features such as DHCP Snooping, IP Source Guard, and ARP security where applicable.

- Enable Spanning Tree Protocol features (for example, BPDU Guard, Loopguard, and Root Guard).
- Use Switch IOS ACLs and Wire-speed ACLs to filter undesirable traffic (IP and non-

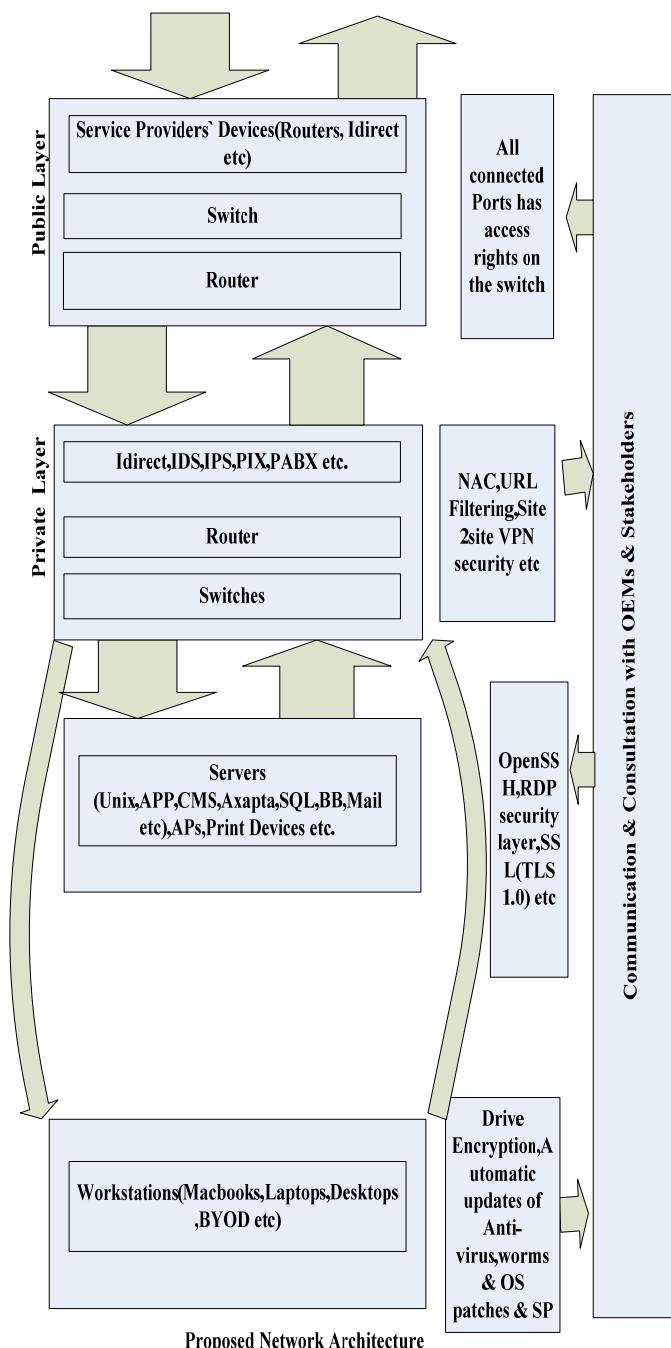


Figure 2: Proposed Layered Network architecture using data and communication devices

This totally eliminates possible entrance of the intentional or accidental attackers/invader. And this drives steady communication and consultation among stakeholders/OEMs to ensure that Operating systems of the devices are always updated both on the private, public and entire corporate network. In addition, necessary advance security measures are implemented at appropriate layers.

With steady communication and consultation of OEMs, the stakeholders are provided with updates, firmwares, patches or service packs and some customized net scanners such as Microsoft Baseline Security Analyzer(MBSA) for windows platform and Open Vulnerability Assessment System(OpenVAS) for open-source related (Cisco, Avaya, android etc.) Platform. These customized applications detect common security misconfigurations and vulnerabilities in the data and communication devices. And with these scanners, organizations can frequently carry out in-depth checks for vulnerabilities and any low priority ones such as patches, OS updates etc. detected are added to the risk register for close watch and monitoring.

The result from this, forms further basis for thorough risk assessment on all the layers of devices to ensure the targeted integrity, availability and confidentiality

VI. CONCLUSION

Every Organization has mission and vision which is backed and driven by strong information and communication technology. And due to ever dynamic and unique nature of information and communication technology, data and communication devices cannot move out way of trouble. In fact, they are more vulnerable to economic and political uncertainties than any other investments. In view of this, above proposed network architecture is key, to ensure that organizations are proactively positioned to eliminate inherent vulnerabilities of common network architecture and take advantage of the proposed solutions.

Furthermore, since attacks assume dynamic forms, it is advisable to charge network Engineers , systems administrators and Experts to make this process a daily routine and carry out aggressive end-users awareness campaign on what to do once they sense data compromise vis-à-vis Integrity, Availability and confidentiality. This is important as Risk assessment is the duty of all stakeholders hence the

encouragement to ensure consistent two-way communication and consultation among stakeholders.

VII. REFERENCE

- [1] Stoneburner G., Goguen, A. & Feringa, A. (2002). Risk Management Guide for Information Systems. Retrieved January 4, 2015 From <http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf>.
- [2] Alshboul, A. (2010).Information Systems Security measures and countermeasures: Protecting Organizational Assets from malicious Attacks, 3, Article 486878. Journals on Communications of the IBIMA, 2010, 1-9. Retrieved March 15 2015 from <http://www.ibimapublishing.com/journals/CIBIMA/2010/486878/486878.pdf>.
- [3] Bhajji, Y.(2008). Security Features on Switches(courtesy Csico press). Retrieved on December 7, 2017 from <http://www.ciscopress.com/articles/article.asp?p=1181682&seqNum=3>
- [4] Chia, T., (2012). Confidentiality, Integrity, Availability: The three components of the CIA Triad. <http://security.blogoverflow.com/2012/08/>

[Confidentiality-integrity-availability-the-three-components-of-the-cia-triad/](http://www.csicopress.com/articles/article.asp?p=1181682&seqNum=3)
- [5] Manes C.(2014). The 21 most common misconfigurations that will come back to haunt you!Retrieved March 20 2015 from <http://www.gfi.com/blog/the-21-most-common-misconfigurations-that-will-come-back-to-haunt-you/>
- [6] Kosutic, D., (2014). Risk Assessment vs. Business Impact Analysis. <http://webcache.googleusercontent.com/search?q=cache:8eF68VJu0cYJ:advisera.com/27001academy/knowledgebase/risk-assessment-vs-business-impact-analysis/+&cd=1&hl=en&ct=clnk&gl=ng>
- [7] Pascucci M.(2012).Network Security Horror Stories: Router Misconfigurations.Retrieved March 22 2015 from <http://blog.algosec.com/2012/09/network-security-horror-stories-router-misconfigurations.html>
- [8] IRS Office of Safeguards Technical Assistance Memorandum Protecting Federal Tax Information (FTI) Through Network Defense-in-Depth. Retrieved April 20, 2015 from <https://www.irs.gov/pub/irs-utl/protecting-fti-throughnetworkdefense-in-depth.doc>.
- [9] Cisco Press(2014). Cisco Networking Academy's Introduction to Basic Switching Concepts and Configuration .

Sag-Tension Analysis of AAAC Overhead Transmission lines for Hilly Areas

Muhammad Zulqarnain Abbasi, M. Aamir Aman, Hamza Umar Afridi, Akhtar Khan

IQRA National University, Pakistan

Abstract—Power system is the transfer of electricity from generation to the point of user location. Power system is composed of generation of power, its transmission and distribution. Transmission system is the main part out of these three in which mostly losses occur. The unchanging factors of the transmission line on which these losses depend are inductance, resistance and capacitance. These constants or unchanging factors play a vital role in the performance of transmission line. For example the capacitance effect will be more and its performance will be affected if the height of transmission line is less from the ground. On the other hand its capacitance will be less but tension will be high if the height of the transmission is high. For this reason a transmission line is connected in a curved or catenary shape known as sag. To minimize tension sag is provided in a transmission line. Sag and tension must be adjusted in safe limits. This immediate paper gives a simulation structure to calculate sag and tension of AAAC (All Aluminum Alloy Conductors) of overhead transmission lines with same span length for minimum operating temperature. Three different cases are presented with different towers height and are explained in detail for unequal level span. The results show that the tension and sag increased with height. So great the height difference, higher tensions upon higher towers.

This paper will be very helpful to find the sag-tension values of AAAC conductor for Unequal level supports without calculating it mathematically.

Keywords-component;sag;tension;transmission system

I. INTRODUCTION

System transferring electricity from generation to transmission is known as transmission system. In transmission system transmission lines and substations play a major role. In power system network the biggest part is incorporated by the lines transporting electricity. Designing and erecting transmission system require proper construction or modeling of these lines. A transmission systems successive execution depends on kind of transmission model used in the system. Catenary is the term given to the curve shape in which transmission lines are connected between supportive towers or power. Transmission line is never connected in straight line. To minimize the tension in the transmission system sag is provided to the transmission line. Similarly, if the tension is high in transmission line, sag

will be minimum and there is a possibility that the conductor may break.

Sag and tension are inversely proportional to each other. However, if sag increase the size of conductor used also increases and the cost is raised. The distance between two towers depends on the sag intensity. If the distance between two towers is significant the sag will also be large.

Sag-Tension computations are aimed at fixing suitable restrictions between sag and tension in order to continue or steady uninterrupted power supply to the users.

With Sag-Tension calculations we can determine the conductor temperature as well as ice and wind; load concurrently. Tension is the limit of towers and conductor to keep the tension limited. Sag clearance distance is dependent on line crossings and ground. If the clearance distance is less than crossing distance chances are that line faults may occur.

The “V” or “I” configuration also play important role for calculating sag-tension along with quantity of insulator strings. Naturally, insulator string possesses the attributes of an element. Being element provides add up of great distance to sag created by the conductor.

Considering bunch of conductors is also important in different cases. For each phase two or more conductors are used. Extra high voltage system may use two bunches of conductors per phase. Occasionally, a substation that is accumulating power from generating station may use three conductors or power lines for each phase. Hence, the foundations regarding the sag-tension approximation process are obligatory to examine and guarantee or certify that the result match the true state.

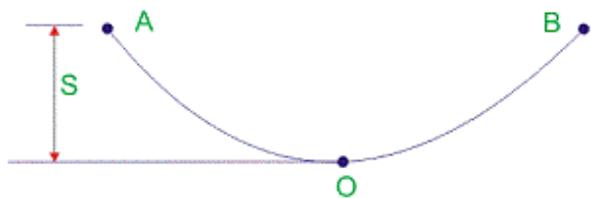


Figure 1. Sag in Overhead Conductor

The above figure is showing that there are two equal supports named as point A and B. while point O is the lowest point between two supports. Similarly "S" is referred as sag which is distance between the point of support and lowest point of conductor.

II. CALCULATION OF SAG

In overhead transmission line designing, it is important to keep sag under safe limit, and at same time tension running in the conductor is also within safe limit. As a matter of fact, tension is administered by weight of conductor, ice load on wires, temperature variation and effect of wind. According to common practice, the tension of conductor is kept under 50% of its ultimate tensile power. It means least factor of safety of a conductor tension needs to be two. Now sag as well as tension calculation of a conductor for Unequal support will be carry out [8].

A. When supports are at Un- equal levels

In hilly areas, we generally come across conductors

Suspended between supports at unequal levels. Figure 2. shows a conductor suspended between two supports A and B which are at different levels. The lowest point on the conductor is O.

Let

l = Span length

h = Difference in levels between two supports

X_1 = Distance of support at lower level (i.e. A) from O

X_2 = Distance of support at higher level (i.e. B) from O

T = Tension in the conductor

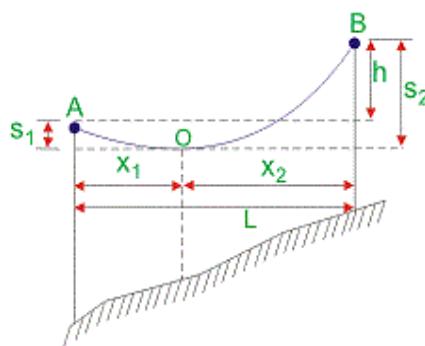


Figure 2. When supports are at Unequal levels

If w is the weight per unit length of the conductor, then

$$\text{Sag } S_1 = \frac{wx_1^2}{2T}$$

and

$$\text{Sag } S_2 = \frac{wx_2^2}{2T}$$

Also

$$x_1 + x_2 = l \quad (i)$$

Now

$$S_2 - S_1 = \frac{w}{2T}(x_2^2 - x_1^2) = \frac{w}{2T}(x_2 + x_1)(x_2 - x_1)$$

$$S_2 - S_1 = \frac{wl}{2T}(x_2 - x_1)$$

But

$$S_2 - S_1 = h$$

$$h = \frac{wl}{2T}(x_2 - x_1)$$

$$x_2 - x_1 = \frac{2Th}{wl} \quad (ii)$$

Solving eq. (i) and (ii), we get

$$x_1 = \frac{l}{2} - \frac{Th}{wl}$$

$$x_2 = \frac{l}{2} + \frac{Th}{wl}$$

III. METHODOLOGY

For the result-oriented sag-tension of AAAC transmission line considering different unequal span heights of minimum operating condition are analyze in this research paper.

The tool used for the calculation is ETAP. The module ETAP containing an analytical strength for Transmission and Distribution Line Sag and Tension calculation. It is easily available low cost simulation software to calculate the appropriate sag & tension in order to ensure appropriate operating conditions on the overhead transmission lines.

The spans length for all cases is same which are 200m and the conductors configuration is set as horizontal.

Three Different cases i.e. 1, 2 & 3 are considered.

- In case 1 calculated sag-tension of AAAC overhead transmission lines under minimum operating condition with the height difference of 10m.
- In case 2 calculated sag-tension of AAAC overhead transmission lines with same temperature as previous one but the height difference is 30m.
- In case 3 calculated sag-tension of AAAC overhead transmission lines under minimum operating condition with maximum height difference of 50m.

These calculations are for unequal level spans only and when both the towers are at height difference of 10m, 30m and 50m respectively.

The conductor used in this paper is AAAC (All Aluminum Alloy Conductor) because:

These conductors are of high strength made of Aluminum-Magnesium-Silicon alloy and are having better ratio of strength to weight enabling the conductors to exhibit more efficient electrical characteristic. They have excellent sag-tension characteristics and superior corrosion resistance when compared with other conductors.

Comparing with traditional ACSR, AAAC are lighter in weight, are having lower electrical losses and comparable strength and current carrying capacity.

IV. RESULTS AND DISCUSSION

A. Case 1

In the 1st case, sag-tension under minimum operating temperature i.e. 5°C with a height difference of 10m is analyzed. As in hilly areas towers are at different heights so in this case we have considered only 10m height difference.

TABLE I. MINIMUM TEMPERATURE WITH 10M HEIGHT DIFFERENCE

Type of Conductor	Minimum Temperature (5 °C) and height (10m)			
	Tension		Sag	
	Low Tower	High Tower	Low Tower	High Tower
AAAC	1697	1707	0	10

In the above table sag and tension results of minimum operating temperature with 10m height difference is shown. As it shows above that when support are at 10m height difference the tension on high tower is 1707 while on low tower it is 1697. Similarly incase of sag high tower has more sag 10m than the lower tower. The results are also shown below with the help of graph.

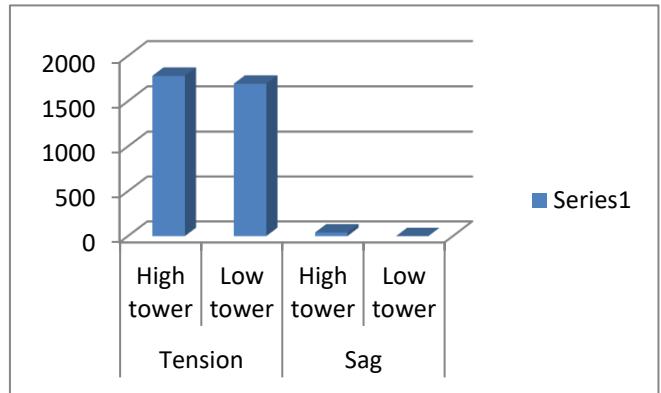


Figure 3. Sag-Tension Results with 10m height difference

B. Case 2

In this case temperature is same as the previous one but the height difference is 30m. As mentioned earlier that hilly areas have different heights for supporting towers so therefore in this case we have consider 30m height difference.

TABLE II. MINIMUM TEMPERATURE WITH 30M HEIGHT DIFFERENCE

Type of Conductor	Minimum Temperature (5 °C) and height (30m)			
	Tension		Sag	
	Low Tower	High Tower	Low Tower	High Tower
AAAC	1711	1738	0	30

As from the above table sag and tension results of minimum operating temperature with 30m height difference are given. When supports are at 30m height difference the tension on high tower is 1738 while on low tower is 1711. Similarly incase of sag high tower has more sag 30m than the lower tower because high tower exerts high tension. The results are also shown below with the help of graph.

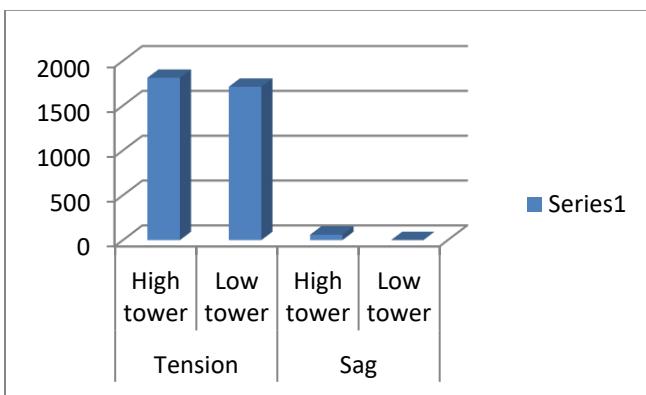


Figure 4. Sag-Tension Results with 30m height difference

C. Case 3

In 3rd case, sag and tension with minimum operating temperature and height difference of 50m is consider.

TABLE III. MINIMUM TEMPERATURE WITH 50M HEIGHT DIFFERENCE

Type of Conductor	Minimum Temperature (5 °C) and height (50m)			
	Tension		Sag	
	Low Tower	High Tower	Low Tower	High Tower
AAAC	1747	1793	0	50

In the above table sag and tension results of minimum operating temperature with 50m height difference is mention. The above table shows that when support are at 50m height difference the tension on high tower is 1793 while on low tower is 1747. Similarly incase of sag high tower has more sag 50m than the lower tower. The results are also shown below with the help of graph.

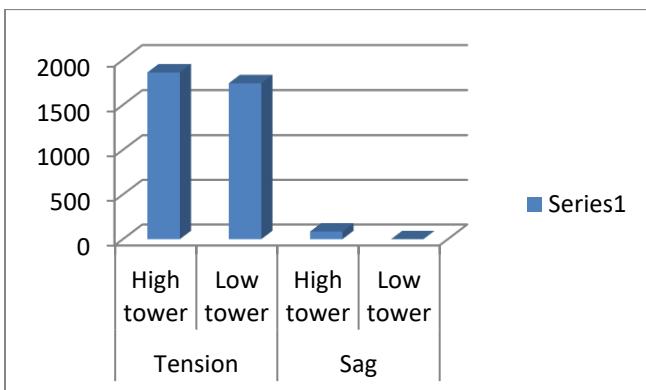


Figure 5. Sag-Tension Results with 50m height differnce

V. CONCLUSION AND FUTURE WORK

This research paper has three different cases which are taken from analysis of sag and tension estimation for AAAC Overhead transmission lines. Equal span length is considered for all cases. Three different towers height were considered with minimum operating conditions. From the results following conclusion is drawn:

In 1st case the temperature is minimum but the height difference was 10m. So when the height difference was minimum (10m) the tension in higher tower is high than the lower tower.

Similarly for 2nd case, the temperature was set same as the previous case but the height difference increased from 10m to 30m. As the height increases the tension will act more force on higher tower as compare to low tower.

In the 3rd case temperature was same again but the height difference was maximum (50m). So due to high difference in towers height the tension on higher tower is also greater due to height. In AAAC the tension and sag increased with height. So great the height difference, higher tensions upon higher towers.

From this paper, one can easily find the sag-tension values of AAAC conductor for Unequal level supports without calculating it mathematically.

In future, the sag-tension estimation of other conductors will also be considered in ETAP.

REFERENCES

- [1] I. Albizu, A. J. Mazon, and E. Fernandez (2011). "A method for the Sag-tension, calculation in electrical overhead lines. International Review of Electrical Engineering, volume 6, No. 3 pp. 1380-1389
- [2] J. Quintana, V. Garza, C. Zamudio(2016). Sag-Tension Calculation Program for Power Substations. IEEE PES Asia-Pacific Power and Energy Conference- Xi'an-China.
- [3] Maamar Taleb, Mohamed Jassim Ditto, Tahar Bouthiba (2006). Performance of Short Transmission Lines Model . University of Bahrain Department of Electrical and Electronics Engineering.
- [4] Oluwajobi F. I., Ale O. S. and Ariyanniuola A (2012). Effect of Sag on Transmission Line. Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 3 (4): 627-630 © Scholar link Research Institute Journals, (ISSN: 2141-7016).
- [5] R.G. Olsen, K.S Edward (2002) "A New method for Real Time Monitoring of High Voltage Transmission line Conductor sag. IEEE Transactions on Power Delivery Vol. 17, No. 4, pp 1142-1152.
- [6] Sag-Tension Calculation Methods for Overhead Lines (2007). CIGRE B2-12 Brochure (Ref. No. 324) pp. 31-43.
- [7] Seppa T O, 1992, A Practical Approach For Increasing the Thermal Capabilities of Transmission lines, IEEE/PES Summer Meeting, pp 1536-1542
- [8] S. Kamboj R. Dahiya (2014) Case Study to Estimate Sag in Overhead Conductors Using GPS to Observe the Effect of Span Length. 978-1-4799-3656-4/14/\$31.00 2014 IEEE
- [9] V.K. Mehta and Rohit Mehta (2014). Principles Power System. S. Chand and Company Pvt. Ltd. Ram Nagar New Delhi.

A Review of Intelligent Agent Systems in Animal Health Care

Omankwu, Obinnaya Chinecherem, Nwagu, Chikezie Kenneth, and Inyama, Hycient

**¹ Computer Science Department, Michael Okpara University of Agriculture, Umudike
Umuahia, Abia State, Nigeria**
saintbeloved@yahoo.com

**² Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
Nwaguchikeziekenneth@hotmail.com**

**³ Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka
Anambra State, Nigeria.**

ABSTRACT

This paper discusses the several research methodologies that can be used in Computer Science (CS) and Information Systems (IS). The research methods vary according to the science domain and project field. However a little of research methodologies can be reasonable for Computer Science and Information System.

Keywords- Computer Science (CS), Information Systems (IS), Research Methodologies.

1. INTRODUCTION

Science is an organized or systematic body of knowledge (Jain, 1997). Science embraces many different domains however those domains are related. Logic and mathematics sciences are the core of all sciences. From there and descending it emerge the natural sciences such as physic, chemistry and biology. In the next come the social sciences.

As (Denning, 2005) reported Computer Science (CS) should not be called as a science. Although CS is definitely a recent discipline, few will still argued that it is not provided with attributes to be qualified as a science. Computer Science has its specificity and has its bases in logic and mathematics.

However CS is transversal to very different domains in science. To understand which research methodologies that can be used in CS and IS we have to understand the differences between CS and IS.

Computer science (CS) characterized as an empirical discipline, in which each new program can be seen as an experiment, the structure and behavior of which can be studied (Allen ,2003). In particular, the field of CS is concerned with a number of different issues seen from a technological Perspective, e.g. theoretical aspects, such as numerical analysis, data Structures and algorithms; how to store and manipulate, the relationship between different pieces of software and techniques and tools for developing software .The field of Information Systems (IS) is concerned with the interaction between social and technological issues (Allen ,2003).

In other words, it is a field which focuses on the actual “link” between the human and social aspects (within an organization or other broader social setting), *and* the hardware, Software and data aspects of information technology (IT). In the next sections we will discuss the different methods of research methodology and its reasonability for the CS and IS domains.

Research Methods

Before we start in discuss the different types of research methodologies we have to define the research. In an academic context, research is used to refer to the activity of a diligent and systematic inquiry or investigation in an area, with the objective of discovering or revising facts, theories, applications etc. The goal is to discover and disseminate new knowledge. There are several methods that can be used in CS and IS in next subsection we will show these methodologies.

Experimental Method

Experimental shows the experiments that will occur in order extract results from real world implementations. Experiments can test the veracity of theories. This method within CS is used in several different fields like artificial neural networks, automating theorem proving, natural languages, analyzing performances and behaviors, etc. It is important to restate that all the experiments and results should be reproducible. Concerning, for example, network environments with several connection resources and users, the experiments are an important methodology Also in CS fields and especially IS fields that take in consideration the Human- Computer Interaction. It is mandatory the usage of experimental approaches. If we use the experimental method in IS field we may need to use some methods or tools in conjunction with the experimental method. These methods or tools used to support and prove the legibility of the developed project. For example if a student wants to implement new social network with new concepts or develop an existing social network, who he can measure the

legibility of his implementation? The answer of this question consists of two parts according to the nature of the project. The technical issue is the first part of the project that can be tested by benchmarks like (availability, reliability, scalability, stability, etc.). The usability of the project is the second part of testing that needs a feedback from the users of the system; the results for the second part can be obtained by the statistical analysis of a questionnaire which is a tool used in conjunction with the experimental method.

Simulation Method

Simulation method used especially in CS because it offers the possibility to investigate systems or regimes that are outside of the experimental domain or the systems that are under invention or construction. Normally complex phenomena that cannot be implemented in laboratories evolution of the universe. Some domains that adopt computer simulation methodologies are sciences such as astronomy, physics or economics; other areas more specialized such as the study of non-linear systems, virtual reality or artificial life also exploit these methodologies. A lot of projects can use the simulation methods, like the study of a new developed network protocol. To test this protocol you have to build a huge network with a lot of expensive network tools, but this network can't be easily achieved. For this reason we can use the simulation method.

Theoretical Method

The theoretical approaches to CS are based on the classical methodology since they are related to logic and mathematics. Some ideas are the existence of conceptual and formal models (data models and algorithms). Since theoretical CS inherits its bases from logic and mathematics, some of the main techniques when dealing with problems are *iteration, recursion and induction*. Theory is important to build methodologies, to develop logic and semantic models and to reason about the programs in order to prove their correctness. Theoretical CS is dedicated to the design and algorithm analysis in order to find solutions or better solutions (performance issues, for example). Encompassing all fields in CS, the theoretical methodologies also tries to define the limits of computation and the computational paradigm. In other words we can say that we can use the theoretical method to model a new system. However the theoretical method can help in finding new mathematical models or theories, but this method still needs other methods to prove the efficiency of the new models or theories. For example when a student need to develop a new classifier in AI by using the mathematical representation and theoretical method, he need to prove the efficiency of this model by using one of the previous methods.

Conclusion

In this paper we try to differentiate between the domains of science and CS and IS to understand the best methods that can be used in CS and IS. Each project in CS or IS have its free nature so the paper give examples of different kinds of projects in CS and IS and the proper research methodologies that can be used in these projects.

References

1. R. K. Jain and H. C. Triandis: Management of Research and Development Organizations: Managing the Unmanageable. John Wiley & Sons, (1997).
2. Gordana Dodig-Crnkovic: Scientific Methods in Computer Science. Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden (2002).
3. Denning Peter J.: Is Computer Science Science?. COMMUNICATIONS OF THE ACM, Vol. 48, No. 4 (2005).
4. Allen Newell, Herbert A. Simon: Computer Science as Empirical Inquiry: Symbols and Search. Communication. ACM 19(3): 113-126(1976).
5. Suprateek Sarker, Allen S. Lee: Using A Positivist Case Research Methodology To Test Three Competing Theories-In-Use Of Business Process Redesign. J. AIS (2001).

Simulated Annealing algorithm for VLSI floorplanning for soft blocks

Rajendra Bahadur Singh

Dept. of Electronics & Comm., School of ICT
Gautam Buddha University,
Greater Noida, INDIA
rbs2006vlsi@gmail.com

Anurag Singh Baghel

Dept. of Computer Science, School of ICT
Gautam Buddha University,
Greater Noida, INDIA
asb@gbu.ac.in

Abstract— In the VLSI physical design, Floorplanning is the very crucial step as it optimizes the chip. The goal of floorplanning is to find a floorplan such that no module overlaps with other, optimize the interconnection between the modules, optimize the area of the floorplan and minimize the dead space. In this Paper, Simulated Annealing (SA) algorithm has been employed to shrink dead space to optimize area and interconnect of VLSI floorplanning problem. Sequence pair representation is employed to perturb the solution. The outcomes received after the application of SA on different benchmark files are compared with the outcomes of different algorithms on same benchmark files and the comparison suggests that the SA gives the better result. SA is effective and promising in VLSI floorplan design. Matlab simulation results show that our approach can give better results and satisfy the fixed-outline and non-overlapping constraints while optimizing circuit performance.

Keywords— VLSI Floorplanning, Simulated Annealing Algorithm, Sequence Pair, Dead Space, etc.

I. INTRODUCTION

Floorplanning[1-3] is an essential design step of VLSI physical design automation[4]. It determines the size, shape, and locations of modules in a chip and as such, it estimates the total chip area, interconnects, and, delays. Computationally, it is an NP-hard problem; researchers have suggested various heuristics and metaheuristic algorithms solve it. A primary research problem in the VLSI floorplanning[12] is its representation. The representation of Floorplan determines the size of search space and the complexity of the transformation between a representation and its corresponding floorplan.

Floorplanning is defined as the process of placing circuit blocks to a given 2D boundary. It is a very crucial step in the VLSI physical design, and the quality of floorplanning significantly affects the successive design steps such as placement and routing. In the early days, floorplanning was tractable since the sizes of chips were limited and designers were able to generate desirable floorplans manually. But in recent years the complexity of design become larger and the number of modules also increased so modern floorplan design becomes impossible manually. Computationally, it is an NP-hard problem [21]; researchers have suggested various heuristics and metaheuristic algorithms solve it. A primary research problem in the VLSI floorplanning is its representation. The representation of floorplan determines the size of search space and the complexity of the transformation between a representation and its corresponding floorplan.

From the complexity point of view, it is an NP-hard problem. The search space increases exponentially with the increase in the number of modules therefore to get an optimum solution is an outdaring task. The quality of floorplanning depends on how it is represented. The representations of floorplans determine the size of the search space and the complexity of transformation between its representation and its corresponding floorplan. There are various representation methods for floorplan such as Bounded Sliced Grid (BSG)[11], Corner Block

List (CBL), Transitive Closure Graph (TCG), B-tree[4], O-tree, Sequence Pair, etc. In this paper, simulation has been done for soft blocks.

The constraints of this issue are two-fold:

- (i) All blocks should be kept into a given 2D boundary.
- (ii) There is no overlap between any two blocks.

The objective is to optimize physical quantities such as area, wirelength, time, noise, voltage, and temperature. These physical quantities have sagacious effects on the properties of a chip. In this paper, we performed the simultaneous minimization of area, wirelength, and timing.

In SA, we use two types of data structures to solve the optimization of floorplanning: interior and constraint graph. The purpose of the interior structure is to store the lower left coordinate of each module so that modules can be placed without any overlapping.

Constraint graph is a representation method, used for floorplan. Here Sequence pair constraint graph is used. This graph represents orders of modules placement

II. PROBLEM STATEMENT

VLSI floorplan is to arrange the modules on a chip, so the inputs for the floorplanning is a set of m rectangular modules $S = \{b_1, b_2, b_3, \dots, b_m\}$ with corresponding areas $\{a_1, a_2, a_3, \dots, a_m\}$. Widths, heights, and areas of the modules are denoted as w_i , h_i , and a_i respectively, $1 \leq i \leq m$. The objective of floorplanning is to arrange the modules in such a way, no two modules overlap each other and the area, wirelength, and other performance indices are optimal.

All modules must be in the rectangular frame and/or square frame. In Figure 1, all modules must be packed without any violation of constraints as mentioned above.

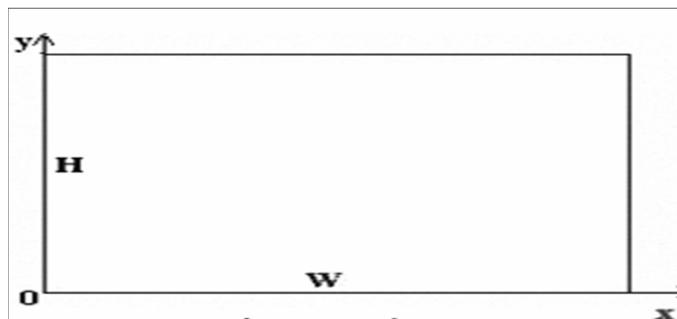


Figure 1. Representation of floorplan

Modules are basically two types, namely hard modules and soft modules. A module is called hard module if its area and an aspect ratio are fixed. A module is called Soft modules if its area is fixed but aspect ratio may vary. The ratio of width and the height of the module are known as the aspect ratio. The floorplan layout can be classified into two types:

A. Sliceable Floorplan

These floorplans can be isolated recursively until each part comprises of a solitary module. A sliceable floorplan is a floor plan that might be characterized recursively as portrayed underneath. Its representation is shown in figure 2..

A floorplan that comprises of a solitary rectangular piece is sliceable.

If a piece from a floorplan can be chopped down in two by a vertical or horizontal line, the subsequent floorplan is sliceable. They can be spoken to by a parallel tree, called the Slicing Tree, having modules at the leaves and cut sorts shown by H (for horizontal cut) and V (for vertical cut) at the internal nodes.

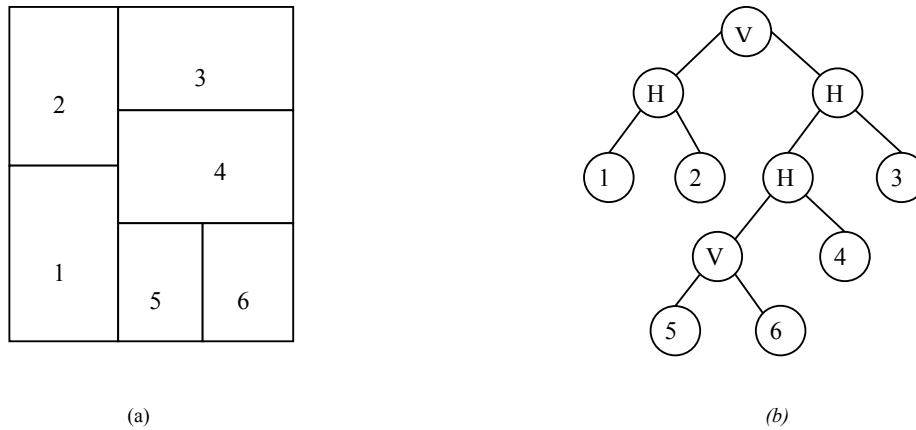


Figure 2. Slicing Floorplan and its tree representation

B. Non-Slicing Floorplan

The non-slicing floorplan can't be cut repetitively. Figure 3, shows a non-slicing floorplan[5] [19].

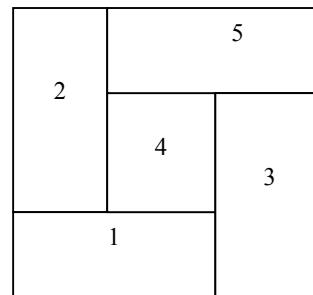


Figure 3. Non-Slicing Floorplan

Floorplanning Cost:

The objective of floorplanning is to arrange the modules in such a way, final output of floorplan structure take the minimal area and minimal interconnect. So the floorplanning cost expression in terms of objective minimal area and minimal interconnect is expresses as

$$Cost(F) = \alpha \left(\frac{Area(F)}{Area^*} \right) + (1 - \alpha) \left(\frac{Wirelength(F)}{Wirelength^*} \right) \quad (1)$$

Where Area (F) is an area of smallest rectangle enclosing all modules. Wirelength(F) is interconnection cost between modules. Area*, and Wirelength* is estimated minimum area, and minimum wirelength respectively. Where α is a weight, $0 \leq \alpha \leq 1$.

The objective of the floorplanning problem is to minimize the cost (F). Minimizing of area is achieved if total dead space is minimized.

Dead Space: It is unused area which is not occupied by any module. The dead space calculation formula is:

$$\text{Dead Space}(D) = \frac{\text{Optimal FP Area} - \text{sum of all modules area}}{\text{Optimal FP Area}} * 100 \quad (2)$$

Optimal FP area is the product of maximum height and maximum width of the floorplan.

Wirelength Estimation: Half-perimeter wirelength is calculated for all the nets to estimate total wirelength required on the chip. Wirelength between two modules can be computed as follows:

$$L = X_{\max} - X_{\min} + Y_{\max} - Y_{\min} \quad (3)$$

Where, X_{\max} and X_{\min} are maximum and minimum value of x-coordinate after placing two modules. Similarly, Y_{\max} and Y_{\min} are the maximum and the minimum of y-coordinate after placing two modules. Thus if floorplan problem has n number of nets then total wirelength on the chip can be computed as

$$\text{Total wirelength} = \sum_{i=1}^n L(i) \quad (4)$$

Problem Description

- Inputs of the problem are a set of modules with fixed geometries and fixed pin positions.
- A set of nets, specifying the interconnections between pins of blocks.
- A set of pins with fixed positions.
- A set of user constraints, e.g., block positions/orientations, critical nets, if any

III. FLOORPLAN REPRESENTATION

A primary research problem in the VLSI floorplanning is its representation. The representation of Floorplan determines the size of search space and the complexity of the transformation between a representation and its corresponding floorplan. Many researchers have suggested many representation[9] schemes for floorplan representation such as Corner Block List, B-Tree, Polish Expression, O-tree, Sequence Pair and etc. In this paper, Sequence pair representation has been used only.

Sequence Pair

The first time concept of Sequence pair was given by Murata[13]. It is the most flexible representation of all the representations. A sequence-pair ($r+$, $r-$) for an arrangement of m modules is a couple of groupings of the m module names. A sequence-pair imposes horizontal/vertical (HV) constraints for every pair of modules. A sequence pair makes it easier to represent the candidate solution of the stochastic algorithm such as genetic algorithm [6-8]. and simulated annealing. To construct a Sequence Pair, first of all, the Constraint Graph Pair i.e. the HCG and VCG have to be created. For instance, ($P=[6 2 5 4 1 3]$, $N=[5 6 1 4 2 3]$) represent sequence pair of a floorplan of the six modules namely: 1, 2, 3, 4, 5, and 6.

Where P and N is the random sequence of number of modules.

Sequence Pair is a succinct representation of non-slicing floorplan[19] of the modules.

Lemma 1: For a given sequence pair if module a and b is present in (P, N):

- 1) Module a is right to module b in floorplan, if a is after b in both P and N sequence.
 - 2) Module a is left to module b in floorplan, if a is before b in both P and N sequence.
 - 3) Module a is above to module b in floorplan, if a is before b in P and is after b in N sequence.
 - 4) Module a is below to module b in floorplan, if a is after b in P and is before b in N sequence.
- where P and N is the random sequence of the number of modules.

Example 1: P = [6 2 5 4 1 3], N = [5 6 1 4 2 3]

Table I

Module No.	Left of	Right of	Below	Above
1	[3]	[5, 6]	[2, 4]	Nil
2	[3]	[6]	Nil	[1, 4, 5]
3	Nil	[1, 2, 4, 5, 6]	Nil	Nil
4	[3]	[5, 6]	[2]	[1]
5	[1, 3, 4]	Nil	[2, 6]	Nil
6	[1, 2, 3, 4]	Nil	Nil	[5]

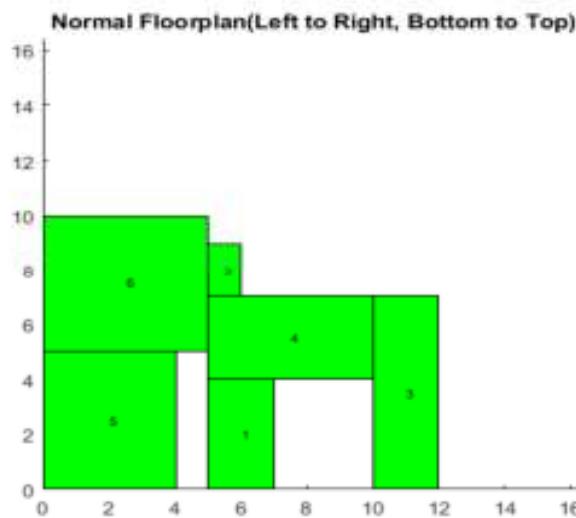


Figure 4. Floorplan representation for example 1 using Sequence pair

Perturb Solution

There are three possible moves in sequence pairs (P, N) to perturb solution to find the better optimal solution.

- 1) Swapping a random pair of modules in the sequence either P or N.
- 2) Swapping the random pair of the modules in both sequences i.e. P and N.
- 3) Rotating a randomly selected module by 90 degrees.

IV. FLOORPLAN ALGORITHMS

As discussed earlier in this paper, floorplanning determines the positions of modules so that objectives like optimization [17] of area and total interconnect can be achieved. We will discuss Simulated Annealing Algorithm here.

Simulated Annealing (SA)

SA [14] is inspired by an analogy between the physical annealing of solids (crystals) and combinatorial enhancement. In the physical annealing process a solid is first melted and then cooled very slowly, spending a long time at low temperatures, to obtain a perfect lattice structure corresponding to a minimum energy state. SA transfers this process to local search algorithms for combinatorial enhancement problems. It does so by associating the set of solutions of the problem attacked with the states of the physical system, the objective function with the physical energy of the solid, and the optimal solutions with the minimum energy states.

Simulated Annealing [4] is widely known algorithm, it can be effectively apply in the VLSI physical design [17] shown in figure 5 and other fields. Before applying it in VLSI floorplanning for optimum results, need to concern four ingredients.

- 1) Solution Space
- 2) Neighborhood structure
- 3) Cost function
- 4) Annealing Structure

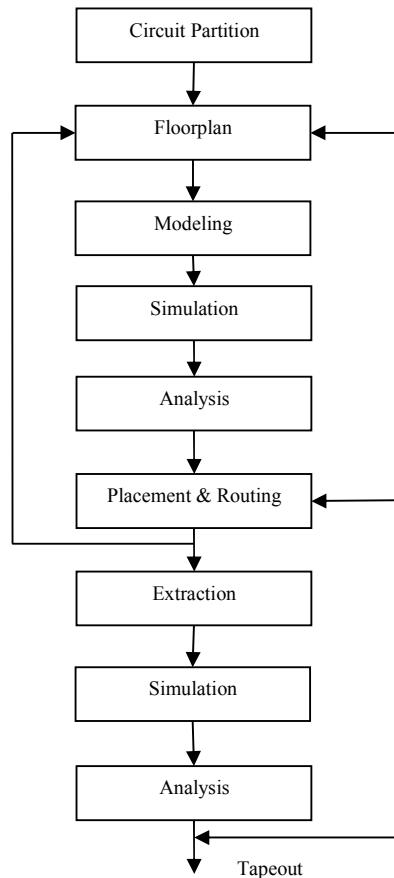


Figure 5. VLSI Physical Design

General algorithmic outline for SA:

Procedure Simulated Annealing Algorithm

1. Input benchmarks circuits.
2. Generate randomly an initial solution.
3. Initialize Annealing parameters such as Temperature, Number of iterations, cooling rate etc.
4. Perturb the initial solution and generate a new solution.
5. If new solution result is better than old solution then replace the old solution with new solution.

Otherwise, accept the new solution with probability $e^{-\Delta C/T}$

6. If the number of iterations at the current temperature reaches the length of Markov chain then switch to next step, otherwise move to step 4.
7. Decrease the temperature by cooling rate(α) i.e,

$$T(i) = \alpha \cdot T(i-1)$$

8. Repeat the step from 4 to 7 until reach to termination criteria.

Where T = initial temperature, Cost (ΔC) = new cost - old cost

SA performs computation that analogous to physical process:

- The energy corresponds to the cost function.
- Molecular movement corresponds to a sequence of moves in the set of feasible solution.
- Temperature corresponds to a control parameter T , which control the acceptance probability for a move i.e. a good move.

V. PERFORMANCE ANALYSIS AND SIMULATION RESULTS

The proposed Simulated Annealing algorithm is implemented in MATLAB programming language, and the experiment was executed on an Intel(R) Core(TM) 2 Duo CPU(2.4GHz, 4GB RAM) machine running windows 2007. The parameters of SA algorithm were set as follows, initial temperature=2000, cooling rate (α) =0.95, number of iterations=20.

The proposed Simulated Annealing algorithm is tested with one of the benchmark circuits named as MCNC (Microelectronic Centre of North Carolina) and find the solutions of modern floorplanning problems [18]. These MCNC benchmarks are standard problems in VLSI floorplanning. The details of MCNC benchmark circuits are shown in table I.

TABLE I. DETAILS OF MCNC BENCHMARK CIRCUITS

S.No.	Problem	Benchmarks	No.of modules	No. of nets	No.of pins
1	Apte	MCNC	9	97	287
2	Xerox	MCNC	10	203	698
3	HP	MCNC	11	83	309
4	Ami33	MCNC	33	123	522

TABLE II. SIMULATON RESULT FOR MCNC BENCHMARKS

Benchmark	Area	HPWL	Dead Space %	Simulation Time(Sec)
Apte	47.35	663754	1.68	35
Xerox	19.66	575710	1.61	49
HP	9.11	227224	3.16	29
Ami33	1.20	101753	3.62	107

TABLE III. COMPARISION OF AREA WITH DIFFERENT ALGORITHM

Benchmark	VOAS[21]	PSO[20]	TCG[10]	SA (our)
Apte	47.1	47.31	46.92	47.35
Xerox	20.3	20.38	19.83	19.66
HP	9.46	--	8.94	9.11
Ami33	1.20	1.29	1.20	1.20

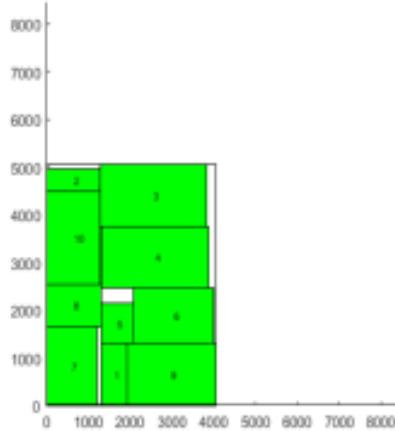


Figure 6. Floorplan Layout for Xerox benchmark

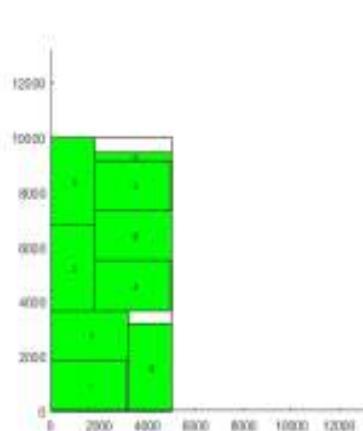


Figure 7. Floorplan Layout for Apte benchmark

Xerox benchmark file contains 10 modules. Using the Simulated Annealing, it takes 20 iterations to improve the performance. Here the total area minimized to 19.66, and the dead space minimized to 1.61%. The Simulated result for “Xerox” file is shown in figure 6.

Apte benchmark file contains 9 modules. Using the Simulated Annealing, it takes 20 iterations to improve the performance. Here the total area minimized to 47.35, and the dead space minimized to 1.6 %. The Simulated result for “Apte” file is shown in figure 7.

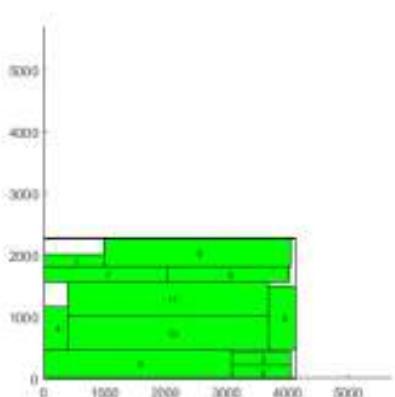


Figure 8. Floorplan Layout for HP benchmark

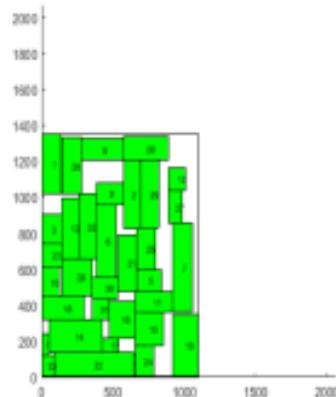


Figure 9. Floorplan Layout for AMI33 benchmark

HP benchmark file contains 11 modules. Using the Simulated Annealing, it takes 20 iterations to improve the performance. Here the total area minimized to 9.11, and the dead space minimized to 3.16. The Simulated result for “HP”file is shown in Figure 8.

Ami33 benchmark file contains 33 modules. After 20 iterations, Simulated Annealing Algorithm has improved the performance. Here the total area minimized to 1.20, and the dead space minimized to 3.62. The Simulated result for “Ami33”file is shown in Figure 9.

VI. CONCLUSION

A Simulated Annealing (SA) algorithm is proposed to tackle VLSI floorplanning problem. In this paper we conclude that optimization of VLSI Floorplanning using simulated annealing with Sequence pair representation gives better result for MCNC benchmarks. From table III, it is obvious that using SA algorithm area optimization is less compare to other algorithms. SA reduces the area but it takes more iteration and more computational time to find optimal solution. Further research on the application of SA in VLSI floorplanning design problem is to minimization of power and thermal optimization can be do. For future work, we have to study more algorithms like Particle Swarm Optimization algorithm, Genetic algorithm [15][16], Ant Colony Optimization[21], Cuckoo search algorithm to solve modern VLSI floorplanning problem with less computational time along with reshaping layout of circuit.

REFERENCES

- [1] Kurbel, Karl, Bernd Schneider, and Kirti Singh. "Solving optimization problems by parallel recombinative simulated annealing on a parallel computer-an application to standard cell placement in VLSI design." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 28.3 (1998): 454-461.
- [2] Nakaya, Shingo, Tetsushi Koide, and Si Wakabayashi. "An adaptive genetic algorithm for VLSI floorplanning based on sequence-pair." *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*. Vol. 3. IEEE, 2000.
- [3] C. J. Alpert and D. P. Mehta, Handbook of algorithm for physical design automation", Auerbach Publications, pp. 139-142, 2008.
- [4] Lichen, Zhu, et al. "An efficient simulated annealing based VLSI floorplanning algorithm for slicing structure." *Computer Science & Service System (CSSS), 2012 International Conference on*. IEEE, 2012.
- [5] Leena Jain and Amarbir Singh, "Non Slicing Floorplan Representations in VLSI Floorplanning: A Summary," *International Journal of Computer Applications (0975 – 8887)*, vol. 71, No. 15, June 2013.
- [6] J. P. Cohoon and W. D. Paris: "Genetic placement," *Proc. IEEE Int. Conf. on CAD*, pp.422425 (1986).
- [7] L. Davis(ed.): "Handbook of Genetic Algorithms," Van Nos- trand Reinhold (1991).
- [8] D. E. Goldberg: "Genetic Algorithms in Search, Optimization, and Machine Learning," Addison-Wesley Publishing Company (1989)
- [9] Laskar, Naushad Manzoor, et al. "A survey on VLSI Floorplanning: Its representation and modern approaches of optimization." *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*. IEEE, 2015.
- [10] Jai -Ming Lin and Yao-Wen Chang, "TCG: A Transitive Closure Graph-based Representation for Non Slicing Floorplan", *Proceedings of Design Automation Conference* , pp. 764-769, 2001.
- [11] S. Nakatake, K. Fujiyoshi, H. Murata and Y. Kajitani, "Module placement on BSG-structure and IC layout applications," *Proceedings of 1996 IEEE/ACM International Conf. on Computer Aided Design*, 1996, pp. 484-491.
- [12] Singh, Rajendra Bahadur, Anurag Singh Baghel, and Ayush Agarwal. "A review on VLSI floorplanning optimization using metaheuristic algorithms." *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*. IEEE, 2016.
- [13] H. Murata, K. Fujiyoshi, and Y.Kajitani, "VLSI module placement based on rectangle-packing by the sequence-pair," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, No. 12, 1996, pp. 1518-1524.
- [14] Chen, Tung-Chieh, and Yao-Wen Chang. "Modern floorplanning based on B/^{*}-tree and fast simulated annealing." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 25.4 (2006): 637-650.
- [15] B. H. Gwee, and M. H. Lim, "A GA with heuristic-based decoder for IC Floorplanning," *Integration, the VLSI journal* 28, 1999, pp. 157-172.
- [16] X. G. Wang, L. S. Yao, and J. R. Gan, "VLSI Floorplanning Method Based on Genetic Algorithms," *Chinese Journal of Semiconductors*, 2002, pp. 330-335.
- [17] D.Jackuline Moni, "Certain Optimization techniques for Floorplanning in VLSI Physical Design," PhD Thesis, Faculty of Information and Communication Engineering, Anna University, June 2009.
- [18] Jai-Ming Lin, and Zhi-Xiong Hung, SKB-Tree: A Fixed-Outline Driven Representation for Modern Floor-planning Problems" *IEEE transactions on VLSI systems*, vol: 20, issue: 3, pp: 473-484, 04 Feb 2011.
- [19] Y. C. Chang, Y. W. Chang, G. M. Wu, and S. W.Wu, "B*-Trees: A new representation for nonslicing floorplans," *Design Automation Conference*, pp. 458-463. 2000.
- [20] Guolong Chen, Wenzhong Guo and Yuzhong Chen, " A PSO -based Intelligent Decision Algorithm for VLSI Floorplanning", *Journal of Soft Computing A Fusion of Foundation, Methodologies and Applications*, Vol. 14, No. 12, pp. 1329 -1337, 2010.
- [21] Chyi Shiang Hoo, Kanesan Jeevan , Velappa Ganapathy and Harikrishnan Ramiah , "Variable Order Ant System for VLSI Multi Objective Floorplanning", *Applied Soft Computing*, Vol. 13, No. 7, pp. 3285-3297, 2013.

A Review of Expert Systems in Agriculture

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem, and Inyama, Hycient

¹ Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
Nwaguchikeziekenneth@hotmail.com

² Computer Science Department, Michael Okpara University of Agriculture, Umudike
Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

³ Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka
Anambra State, Nigeria.

ABSTRACT

Role of expert systems in agriculture and its applications in efficient crop production and protection technologies has been reviewed and discussed in this paper. Different domains of agriculture are highlighted where expert systems can play an important role for an expert in transferring expert-driven knowledge instantly at the level of farmer's field. This paper explores structure of an expert system, role of expert system in agriculture along with details of expert system developed in the different field of agriculture and also possibilities of designing, developing and implementation of an expert system for agriculture would motivate scientists and extension workers to investigate possible applications for the benefit of entire agricultural community.

Keywords: *Expert System, Knowledge base, Inference engine, crop Management, Crop Disease Diagnostic Domain.*

INTRODUCTION

This paper explores the possibilities of designing, developing and implementation of an Expert System for different activities of agriculture in integrated approach. The overall crop production management problems involve, among many others, management of diseases and insect-pests, integrated water and fertilizer managements, crop economics etc. The management problems also include the lack of enough experts and availability of experts at the farmer's field to support the crop growers. Each crop requires entirely different management practices and cropping pattern. Farmers may not know about all the information on production technology, so they need rapid access to all the possible information and need to take fast decisions to manage their crops efficiently and effectively. In order to raise a successful pulse crop and remain competitive, the modern farmers often rely on crop production specialists to assist them in arriving at the timely decision. Unfortunately, crop specialists are not always available for consultation at the nick of the time. To solve this problem, an Expert System (ES) may become a powerful tool which is a dire need of the day for

farmers, extension workers and Government officials. ES can provide on-line information on different crop management issues like diagnosing and controlling noxious and commonly found insect-pests and diseases, crop economics and designing schedule for irrigation and fertilization application etc. This paper explores the possibilities of designing, developing and implementation of an Expert System for different activities of agriculture in integrated approach.

Structure of Expert System

An Expert system can be viewed as having two environments [10, 16]: the system development environment in which the ES is constructed and the consultation environment which describes how advice is rendered to the users. The development environment starts with the knowledge engineer acquiring the knowledge from the expert. This acquired knowledge is then programmed in the knowledge base as facts about the subject and knowledge relationship. The consultation environment involves the user, who starts the process by acquiring advice from the ES. The ES provides a conclusion and explanation, using its inference engine. It is used by end-users (*i.e.* farmers/extension workers in agriculture domain) to obtain expert's knowledge and advice. The three major components that appear in virtually every expert system are the knowledge base, inference engine, and user interface [7,9].

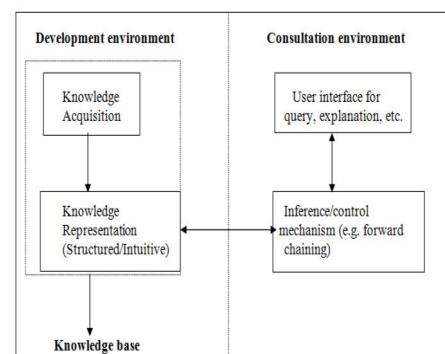


Figure 1: Structure of an Expert System

ROLE OF AN EXPERT SYSTEM IN AGRICULTURE

Traditionally, if a farmer is interested in crop cultivation, he encounters following problems related to information transfer.

A. Static Information:

Information on pulse production technology stored and available in the problem domain reveals that the information is static and may not change according to the growers need. All extension literatures just give general recommendations without taking into consideration all the factors. Hence, generally the information is incomplete.

B. Integration of specialties:

Most of the extension documents deal with problems related to certain specialties: plant pathology, entomology, nutrition, production, etc. In real situations the problem may be due to more than one cause, and may need the integration of the knowledge beyond the information included in different extension documents and books. Image may need, sometimes, an expert to combine other factors to reach an accurate diagnosis, and even if a diagnosis is reached, the treatment of the diagnosed disorder should be provided through extension document.

C. Updating:

Over a period of time the extension documents become obsolete which need to be updated time to time. Changes in chemicals, their doses, and their effect on the environment should be considered while adopting the production and protection technology. Updating this information in the documents and dissemination of the same takes a long time.

D. Information unavailability:

Certain information may not be available in any form of media. It is available only from human experts, extensionists, and/or experienced farmers. In addition, the information transfer from specialists and scientists to extensionists and farmers represents a bottleneck for the development of crop production technology on the national level.

The problems identified can be solved easily using Expert Systems technology in agriculture by using its knowledge base and reasoning mechanism through information acquired from human experts and other sources. Apart from these, the Expert System helps farmers and extension workers to generate all the relevant information and assists them in making environmental friendly and economically viable crop management decisions. Expert system in agriculture, therefore, offers to provide a necessary link between both research information and human expertise and the practical implementation of the knowledge.

EXPERT SYSTEMS IN AGRICULTURE

Expert system evolved as the first commercial product of Artificial Intelligence and is now available in the large number of areas. The potency, scope and appropriateness of expert system in the area of agriculture have been well realized two decades back in developed countries [1,11] and several successful systems have been developed in the field of agriculture. In the recent years, main focus of agricultural expert system applications research is on crop management and plant disease and insect-pests diagnosis [2,3,12,14].

These areas are followed by irrigation management, fertilization management, varietal selection, farm management, crop economics etc. Based on our literature survey, overview of some of the expert systems developed in agriculture fields especially in domain of crop management and crop disease/insect-pest diagnosis and control are presented here.

A. Crop Management Expert Systems

This category of ES includes advisory systems that emphasize the management of specific crops. These systems generally attempt to provide a complete and integrated decision support approach that includes most aspects of growing the crop. There are also crop management advisory systems that focus on specific management issues common to most cropping systems and can therefore be used on a wide range of crops within specific geographic regions[6]. The following ES deals with the growing of a specific crop: GRAPE is an ES for viticulture in Pennsylvania. This ES was developed at Pennsylvania State University in association with Texas A&M University to address the advisory needs of grape growers. This system provides grape growers with recommendations regarding pest management (insect, disease and weed control), fertilization, pruning and site selection. The development environment included the rule based shell called Rule master on the Macintosh microcomputer.

ESIM an Expert System for Irrigation Management was developed for making decisions on water management in an irrigation project. The system was applied to an irrigation management problem of the Mae-Taeng Irrigation Project located in Northern Thailand. The system developed was interactive and made user friendly.

CROPLOT is a rule-based expert system for determining the suitability of crops to given plots. Its use was in the process of plot allocation when planning the production of field crops on the individual farm, usually under severe uncertainties. The system's performance was found satisfactory and a comparison between recommendations of human experts and CROPLOT showed 90% agreement.

COMAX provides information on integrated crop management in cotton. It is designed for use by farmers, farm managers, and county and oil conservation agents. The system uses a combination of expert-derived rules and result generated by the cotton – crop simulation model GOSSYM. This was the first integration of an expert system with simulation model for daily use in farm management.

CUPTEX is an Expert System for cucumber production management under plastic tunnel. System currently provides services on disorder diagnosis, disorder treatment, irrigation scheduling, fertilization scheduling, and plant care subsystems used by agricultural extension of agriculture, and by the private sector. It was developed in KADS. KADS is a methodology for building knowledge-based systems (reference). KADS was used for representation of the interface and task knowledge. Finally, LEVEL 5 object was object for the implementation. It was also the first deployed Expert System in developing countries not only in agriculture but in other fields as well.

CITEX is an Expert System for orange production. It provides services on assessment of farm, irrigation and fertilization scheduling, disorder diagnosis and disorder treatment. System was developed by the Central Laboratory of Agricultural Expert Systems (CLAES), Egypt.

NEPER WHEAT is an Expert System for irrigated wheat management developed at the Central Laboratory of Agricultural Expert System (CLAES) is Egypt. It performs various functions *viz.*, Advice the farmers on field preparation, control pests and weeds, manage harvests, prevent malnutrition, design schedule for irrigation and fertilization, select the appropriate variety for a specific field, diagnose disorder, suggest treatments etc. It is an easy-to-use in Microsoft Windows based application with an English and Arabic interface.

LIMEX is an integrated Expert System with multimedia that has been developed to assist lime growers and extension agents in the cultivation of lime for the purpose of improving their yield. The scope of LIMEX expert system includes: assessment, irrigation, fertilization and pest control. This system was augmented with multimedia capabilities as enhancing an expert system by the integration an expert image, sound, video and data, allows for a good feedback from users, assists in better understanding of the system, and allows for more flexibility in the interactive use of the system. It was developed using an adapted KADS methodology for the knowledge part. CLIPS TM Ver. 6.0 shell on Windows and CLIPS object – oriented language (COOL) was used for development of Expert System.

7.Crop Disease/Insect-pest Diagnostic Expert Systems

Expert System is not new for crop disease diagnostic domain. The weather-based computerized disease forecasters initially developed in the 1970's. Blitecast and the Apple Scab predictive system are examples of forecasters that are currently used to help farmers make decisions about the management of potato late blight and apple scab, respectively [8]. They are similar to expert systems in that the rationale behind their development and application is to aid in the implementation of economically and environmentally sound control practices for particular disease [5]. Of all the systems reviewed, disease/insect-pest management ES were by far the most common, numerous, and widespread. These systems provide farmers, researchers and extension workers with integrated disease and pest management strategies which include all relevant factors in order to adequately and cost effectively control diseases and insect-pests [13]. This requires that many factors such as population dynamics, weather, cost, fungicide and pesticide susceptibility, and the environment be considered in order to reach optimal decisions. Based on our literature survey, we have to mention here some of the expert systems developed for diagnosing diseases of many crops [4, 15] The following Expert Systems are the examples of such crop disease/insect-pest diagnostic systems.

POMME is an Expert system for Apple Pest Orchard Management. It was developed in Virginia to help in managing diseases and insect-pests on apples. This system provides growers with knowledge about fungicides, insecticides, freeze, frost and drought damage, non-chemical care options as well as information from a disease model. External information such as weather data including forecasts and crop symptoms are utilized by the system to generate management decision

recommendations. POMME was one of the first expert systems to incorporate the decision making process of the expert to advise producers in making disease management decisions. The system contains more than 550 rules. PROLOG language was used to build POMME.

VEGES is a multilingual expert system for the diagnosis of pests, disease and nutritional disorder of six greenhouse vegetable *viz.*, pepper, lettuce, cucumber, bean, tomato, and aborigine. It provides the user with a diagnosis on the basis of a brief description of the external appearance of the affected plant. It then suggests method to remedy the problem (e.g., fertilizer, adjustment, fungicides or pesticide applications). The system is accompanied by a new language translation module which allows a non-specialist user (e.g. extension officer) to translate the knowledge base to the native language or dialect of the local farmers.

CPEST is an expert system for managing pests and diseases of coffee in a developing country. Graphical user interface incorporated in CPEST that not only help the farmer in inputting information but also give them visual clues, such as pictures of a pest at different stages of development. System consisted rule-set of 150 production rules. CPEST was built in wxCLIPS, a high-level programming language for constructing expert systems. CPEST's graphical user interface consists of about 40 screens.

AMRAPALIKA is an Expert System for the diagnosis of pests, diseases, and disorders in Indian Mango. The system makes diagnosis on the basis of response/responses of the user made against queries related to particular disease symptoms. A rule-based expert system is developed using Expert System Shell for Text Animation (ESTA). The knowledge base of the system contains knowledge about symptoms and remedies of 14 diseases of Indian mango tree appearing during fruiting season and non-fruiting season.

POMI is an Expert System for integrated pest management of apple orchards. This system was developed cooperatively at the Istituto per la Ricerca Scientifica e Technologica and the Istituto Agrario S. Michele in Italy.

The system provides the apple producer with help on first classifying observations and then providing recommendations on appropriate actions. The KEE development environment was used to construct this system. The system consists of two parts: classification of user findings, and explanation of these findings using abductive reasoning.

CALEX is an Expert System for the diagnosis of peach and nectarine disorders. System diagnoses 120 disorders of peaches in California, including insects, diseases, and cultural problems. The user begins a session by identifying a area on the plant where the problem occurs. The Expert System uses "Certainty Factor" to arrive at conclusions. At the end of a session the Expert System displays all conclusions reached with corresponding levels of certainty.

DDIS is a distance diagnostic and identification system developed at the University of Florida. The system allows users to submit digital samples obtained in the field for rapid diagnosis and identification of plants, diseases, insects, and animals. DDIS provides an environment for agricultural extension agents and specialists to share information on plants, insects and diseases. DDIS is a Java-based three-tier application using Java Remote Method Invocation (RMI) and object database technologies. The system creates a digital image library with associated site, crop, and pest or disorder data that could be used in educational program, assisted diagnosis, and data mining.

D-CAS is an expert system for aid in the appraisal and treatment of diseases of sugarcane. This multimedia computer program was designed and used with Windows as an expert system for identification and control of 59 sugarcane diseases. The program comprises 3 modules: diagnosis, data sheets with pathogen characteristics (including geographical distribution, symptoms, strains, transmission, host range, conditions favorable for disease development, economic importance and control) and data on diseases recorded in 19 parts of the world.

CONCLUSION

The theory of Expert System is well developed & matured and can be applied to a wide spectrum of business problem. Perhaps one of the greatest hindrances in increasing crop production today is that of transferring new agriculture technologies developed at laboratories to the farmer's field. Expert System technology is an ideal approach for transferring the crop production technologies to the farmer's level, the ultimate consumer of agriculture research. Expert systems are not static but dynamic devices as there is always scope for improvement and up gradation.

The approach for the development of Expert System is not difficult to understand and tools for developing expert system are readily available, even some are freely available on internet. The careful development and logical use of Expert System in agriculture can help bridge the gap between research worker and extension worker.

This view of the future is the result of studies and experience gained through the Expert System currently being developed and implemented. Therefore, "Expert System in Agriculture" aims to train extension workers and distribute the Expert System to all extension sites nation-wide. Taking the reducing prices of computers/mobiles into consideration, internet connectivity at village level or panchayat level for the said purpose can be achieved.

REFERENCES

- 1.Carrascal, M.J. and Pau, L.F., A survey of expert systems in agriculture and food processing. *AI Applications*, 6(1992), pp. 27-49.
2. Krause, R.A., Massie, L.B. and Hyre, R.A., Blitecast: A computerized forecast of potato late blight, *Plant Disease Reporter*, 59(1975), pp. 95-98.
3. Kozai, T. and Hoshi, T., Intelligent Information System for Production Management in Agriculture and Horticulture, *Future Generation Systems*, 5(1989), pp. 131-136.
4. Kolhe, S. and Gupta, G. K. , Web-based Soybean Disease Diagnosis and Management System, *Fifth Conference of the Asian Federation for Information Technology in Agriculture (AFITA)*, 2006), pp. 553-559.
5. Chakraborty, P. and Chakrabarti, D.K., A brief survey of computerized expert systems for crop protection being used in India, *Progress in Natural Science*, 18(2008), pp. 469-473.
6. Chu YunChiang, Chen TenHong, Chu-YC, and Chen-TH, Building of an expert system for diagnosis and consultation of citrus diseases and pests, *Journal of Agriculture and Forestry*, 48(1999), pp. 39-53.
- 7..Devraj and Jain Renu, PulsExpert: An expert system for the diagnosis and control of diseases in pulse crops, *Expert Systems with Applications: An International Journal*, 38(2011), pp. 11453-11471.
8. Kramers, M.A., Conijn, C.G.M. and Bastiaansen, C., EXSYS, an Expert System for Diagnosing Flowerbulb Diseases, Pests and Non-parasitic Disorders, *Agricultural Systems*, 58(1998), pp. 57-85.
- 9.Luger, G.F., Artificial Intelligence: Structures and Strategies for Complex Problem Solving, Pearson Education, Inc. (Singapore) Pte. Ltd. (2002).
- 10.Patterson, D.W., Introduction to Artificial Intelligence and Expert Systems, Prentice-Hall of India Pvt. Ltd.(2004).
- 11.Perini, A. and Susi, A., AI in support of Plant Disease Management, *AI Communications*, 18(2005), pp. 281-291.
- 12Plant, R.E., Zalom, F.G., Young, J.A. and Rice, R.E., CALEX/peaches, an expert system for the diagnosis of peach and nectarine disorders, *Horticulture Science*, 24(1989), pp. 700.
- 13.Potter WD, Deng X., Li J., Xu M., Wei Y., Lappas I., Twery MJ, Bennett DJ and Rauscher HM, A web-based expert system for gypsy moth risk assessment, *Computers and Electronics in Agriculture*, 27(2000), pp. 95-105.
- 14.Robinson, B., Expert Systems in Agriculture and Long-term research, *Canadian Journal of Plant Science*, 76(1996), pp. 611-617.
- 15.Saunders, M., Haeseler, C., Travis, J., Miller, B., Coulson, R., Loh, K., et al., GRAPES: an expert system for viticulture in Pennsylvania, *Artificial Intelligence Applications*, 1(1987), pp. 13-20.
- 16.Turban, E. and Aronson, J.E., Decision Support Systems and Intelligent Systems, Pearson Education Asia (2002).

Keywords- Based on Arabic Information Retrieval Using Light Stemmer

Mohammad Khaled A. Al-Maghasbeh¹, Mohd Pouzi Bin Hamzah²,

^{1,2}-School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

Abstract

Arabic morphology complex is a core challenge of Arabic information retrieval. This limitation makes Arabic language is so complex environment of information retrieval specialists due to the high inflected in Arabic language. This paper explains the importance of the stemming in information retrieval through developed a method to search about

Introduction

There are many researches have been conducted in information retrieval, question answering, Arabic morphology, and light stemming due to Arabic language represents inflectional and derivational language [7]. Identify the original of word needs to remove prefix, suffix, and other connected character. This process is known as a stemming task. Arabic language has a complexity morphology that makes NLP applications in Arabic language such as information retrieval is very difficult [8]

In addition, stemming is a morphology process to reduce the derived forms of the word. It helps to find the basic or origin of the specific word (root). The root of specific word is the origin of word. Root, also is a basic characters of word without suffix, and prefix. Light Stemming is a process to derivate the roots of each word that written in natural language (NL) such as Arabic text. In other words, it represents that task to generate the morphological form of the word by removing the all diacritics, suffix, and prefix [3][4].

This paper is organized as follows. Section 2 briefly describes the related works. Section 3 describes information retrieval. Section 4 about proposed system. Section 5 discuss and experimental results. In section 6 it summarizes of the work and future work.

information needs in Arabic text. The new developed method based on light stemmer to increase matching the keywords between the query, with the related document in the test collection.

Keywords

Arabic Information retrieval, keyword-searching, light stemmer, stemming, Arabic-morphology, natural languages processing.

2. Related work

The Arabic language is one of the complex languages in the world. Therefore, it needs to special tools and models to deal with its morphology. In study that conducted by Elabd, E., et al, a new approach is applied to deal with information retrieval in Arabic language through analyze all challenges that face the most current used method in information retrieval systems such as latent semantic indexing (LSI), Latent analysis indexing (LAI), Boolean model, and others for attempting to solve them. In the developed approach, also, the query was processed via divide it into two type; the first one is the query that contains only one word, and second one it contains multiple words. In the first case, the query has been stemmed in preparation to match with all documents which consists this word.

However, on the other side of type of query, the stop-words was removed from the phrase, after that stemming all words to match them with related synonyms with all documents that contains at least one of these words [5]. Samy. et al in their paper indicated medical words extraction in multilingual medical resources as a case study of information extraction. The extraction operation has been done by taking the newswire in health field as a dataset sample to attempt compared them with medical list of common Arabic word in Latin prefix and suffix, where as a several tests

were conducted by applying several samples, after that, the results were good [10]. Al-Taani. et al presented a new method to parse the Arabic simple sentences using a context free grammar (CFGs) for attempting to remove the ambiguity of Arabic language grammars and to enrich researches of the natural language processing (NLP) field with a computation system. This method carried out to test 70 of different simple Arabic statements through converts the statements and words into production rules, whereas the results were very good for all tested statements [4].

Abdelali. et al built a project by using cross-languages information retrieval (CLIR) approach that represent a bilingual method. That method aimed to match between the language of the query and the language of the target to facilitate the desired retrieval [1]. Haav. et al, proposed method known as keyword- based information retrieval, whereas these methods currently are extensively used in web search engine. As a result, this type of information retrieval method has a lack for retrieving and fetching all relative information [6].

The common Arabic search engine depends on the keyword in searching about the answer of user's query. Using the web semantic improved the search operation through adds a new layer into the current web that contains a Meta data (or data about data) to related some concepts with each other. Al-khalifa. et al., focused in their search of analyzing the semantic relation among concepts using intelligent characteristics; but there are some challenges in web search understandability of all complex concept, and deep knowledge inside the Arabic textual documents

4. Proposed system

The proposed system contains several phase. It starts from preprocessing phase that include normalization tokenization, and stop-word removal. After that using light stemming the terms in both documents, and query. The next step index the document keywords, finally matching between the query-keywords with indexed-terms. These phase will be in briefly explain as follow.

[2]. Study of Noordin. et al., presented a project that designed to retrieve information and versus from Holy Quran to facilitate discover and acquires the knowledge from the Quranic texts [9].

In study of Vallet. et al, a new method was proposed to search about some document in web by using an ontology – based knowledge base to increase and improve the accuracy of search. It's being done through used the related concepts, and synonyms to represent the knowledge, and the build the knowledge base to facilitate of retrieval the correspond documents of queries [11].

3. Information retrieval

The information retrieval includes a lot of methods that used of retrieval. Some of these methods depend on the keywords in search about document or any information through match the query-keywords with the relevant information [6]. So improves the performance of information retrieval systems represents an important task for the majority of the information retrieval researchers [11]. It has recently spread too many studies about the passage retrieval as a branch of information retrieval fields. The passage retrieval is one of information retrieval tasks, which it refers to retrieve a portion of the document. There are some passage retrieval methods such as support vector machine, mixture of language model, and other, which majority of them deals with the frequency or density of required word in each passage to compute the relative ranking of relevant passages [12].

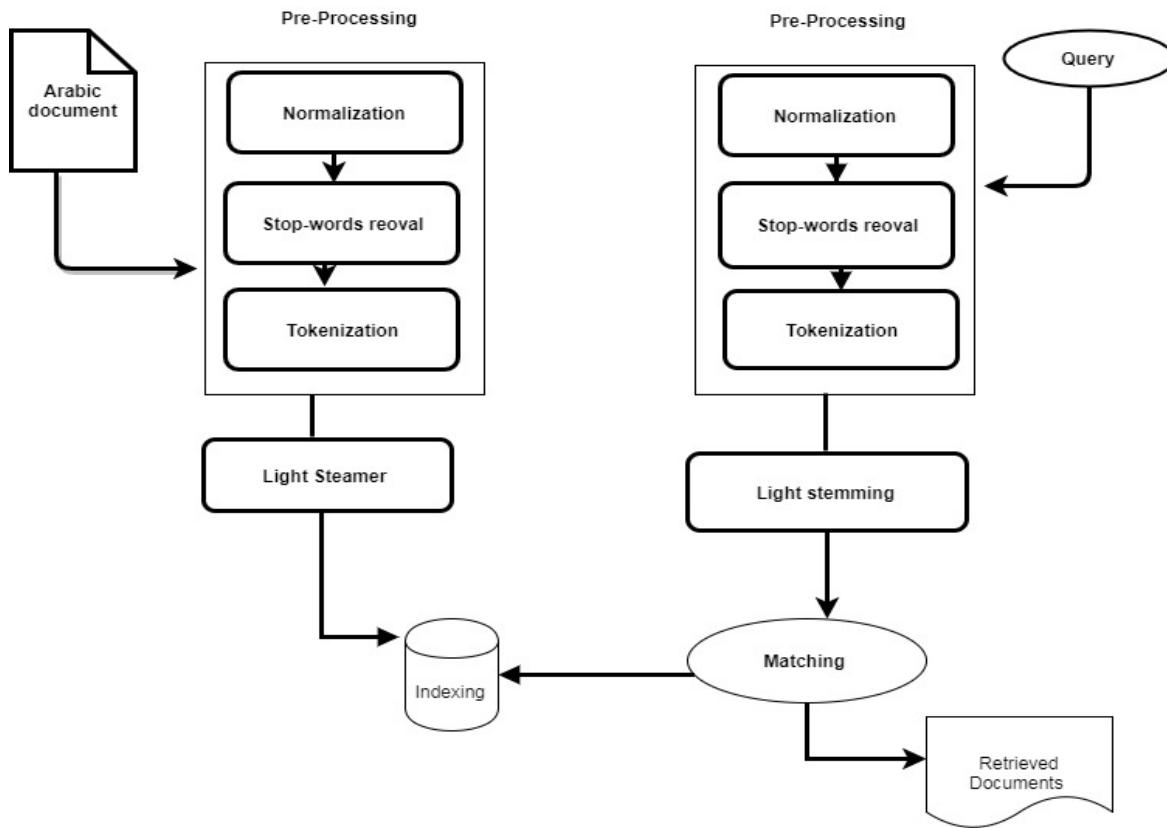


Figure 1: Proposed system architecture

4.1 System phases

4.1.1 Preprocessing phase

This phase is an important process in both, document, and query. The input of this phase is text of modern standard language of Arabic Newswire. It used to reduce the noise in the texts, through remove irrelevant or not important words such as stop words, prepositions, punctuation marks, digits from Arabic texts. After that, replace some Arabic characters into other characters to be more understandable and readable by computer.

Text normalization is applying on several natural language texts. It represents a task to transfer the inconsistency text to be more consistency. In the Arabic language was used normalization to remove the diacritics marks, and normalize the other specific characters. **Tokenization** is a process to divide the plain text into tokens to remove the noise from the text. After that sent it into morphological analyzer to continue the processing [3]. **Stop-words removal** process is to remove the frequent Arabic words that insignificant words or aren't carry important meaning.

6. Discussion

The proposed approach has been applied on a sample of 40- documents, and 3- queries. The Figure is shown below contains the related top documents that retrieved using both proposed method with light-stemmer, and without stemming.

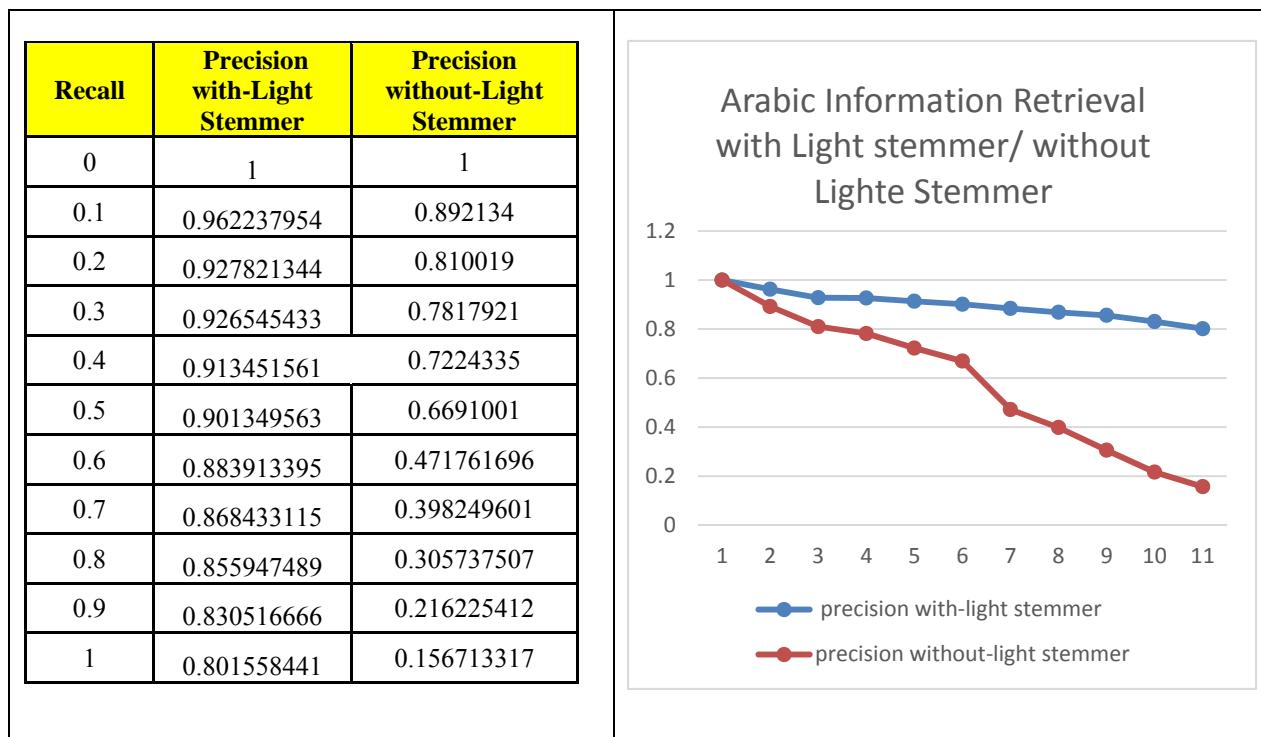


Figure 1: The performance of Arabic retrieval using Light stemmer, and without stemmer

The results of this proposed method using light-stemmer show a good measurement better than without stemming. The interpolated precision to explain of improvement of the performance Arabic information retrieval using the proposed method.

7. Conclusion

This study based keywords- searching through light stemmer to retrieve the information needs from spread Arabic contents in the web. It shows the effectiveness of light stemming on Arabic information retrieval. Light stemming helps the retrieval systems, and search engines through enhancing the search performance of these systems. This study confirms that the light stemming helps to match many of the words that share in the origin of the word or root by deleting the prefix and suffixes, thus allowing the matching of the most different words in the text.

References

- Abdelali, A., Cowie, J., Farwell, D., Ogden, B., & Helmreich, S. (2003). *Cross-language information retrieval using ontology*. Paper presented at the Proc. of the Conference TALN 2003.
- Al-Khalifa, H., & Al-Wabil, A. (2007). *The Arabic language and the semantic web: Challenges and opportunities*. Paper presented at the The 1st int. symposium on computer and Arabic language.
- Al-Taani, A. T., & Al-Rub, S. A. (2009). A rule-based approach for tagging non-vocalized Arabic words. *Int. Arab J. Inf. Technol.*, 6(3), 320-328.
- Al-Taani, A. T., Msallam, M. M., & Wedian, S. A. (2012). A top-down chart parser for analyzing arabic sentences. *Int. Arab J. Inf. Technol.*, 9(2), 109-116.
- Elabd, E., Alshari, E., & Abdulkader, H. (2015). Semantic Boolean Arabic Information Retrieval. *arXiv preprint arXiv:1512.03167*.
- Haav, H.-M., & Lubi, T.-L. (2001). *A survey of concept-based information retrieval tools on the web*. Paper presented at the Proceedings of the 5th East-European Conference ADBIS.
- Hammo, B., Abu-Salem, H., & Lytinen, S. (2002). *QARAB: A question answering*

- system to support the Arabic language.* Paper presented at the Proceedings of the ACL-02 workshop on Computational approaches to semitic languages.
- 8. Larkey, L., Ballesteros, L., & Connell, M. (2007). Light stemming for Arabic information retrieval. *Arabic computational morphology*, 221-243.
 - 9. Noordin, M. F., & Othman, R. (2006). *An information retrieval system for Quranic texts: a proposed system design.* Paper presented at the Information and Communication Technologies, 2006. ICTTA'06. 2nd.
 - 10. Samy, D., Moreno-Sandoval, A., Bueno-Díaz, C., Salazar, M. G., & Guirao, J. M. (2012). *Medical Term Extraction in an Arabic Medical Corpus.* Paper presented at the LREC.
 - 11. Vallet, D., Fernández, M., & Castells, P. (2005). *An ontology-based information retrieval model.* Paper presented at the European Semantic Web Conference.
 - 12. Wan, R., Anh, V. N., & Mamitsuka, H. (2007). *Passage Retrieval with Vector Space and Query-Level Aspect Models.* Paper presented at the TREC.

Kinect Sensor based Indian Sign Language Detection with Voice Extraction

Shubham Juneja¹, Chhaya Chandra², P.D Mahapatra³, Siddhi Sathe⁴, Niles B. Bahadure⁵ and Sankalp Verma⁶

¹Material Science Program, M.Tech first year student, Indian Institute of Technology
Kanpur, UttarPradesh, India

junejashubh@gmail.com

²B.E, Electrical and Electronics Engineering, Bhilai Institute of Technology
Raipur, Chhattisgarh, India

chhava.chandra02@gmail.com

³B.E, Electrical and Electronics Engineering, Bhilai Institute of Technology
Raipur, Chhattisgarh, India

pushpadas27495@gmail.com

⁴Department of Electrical & Electronics Engineering, B.E. Final Year Student, Bhilai Institute of Technology
Raipur, Chhattisgarh, India

sathesiddhi1996@gmail.com

⁵Assoc. Professor at Department of Electronics & Telecommunication Engineering, MIT College of Railway Engineering & Research
Solapur, Maharashtra, India

nbahadure@gmail.com

⁶Assoc. Professor at Department of Electrical & Electronics Engineering, Bhilai Institute of Technology
Raipur, Chhattisgarh, India

sankalpverma99@gmail.com

Abstract

We are progressing towards new discoveries and inventions in the field of science and technology, but unfortunately, very rare inventions could have helped the problems faced by the physically challenged people who face difficulties in communicating with normal people as they use sign language as their prime medium for communication. Mostly, the sign languages are not understood by the common people. Studies say that many research works have been done to eliminate such kind of communication barrier. But those work involves the functioning of Microcontrollers or by some other complicated techniques. Our study advances this process by using the Kinect sensor. Kinect sensor is a highly sensitive motion sensing device with many other applications. Our workflow from capturing of an image of the body to conversion into the skeletal image and from image processing to feature extraction of the detected image hence getting an output along with its meaning and voice. The experimental results of our proposed algorithm are also very promising with an accuracy of 94.5%.

Keywords: Hidden Markov Model (HMM), Image Processing, Kinect Sensor, Skeletal Image

1. Introduction

Since a very long time, we are experiencing a better life due to the existence of various electronic

systems and the sensing elements almost in every field. Physically challenged people find it easier to communicate with each other and common people using different sets of hand gestures and body movements. We hereby provide an aid to very efficiently express themselves in front of common people wherein their sign languages will be automatically converted into text and speeches. Their hand and body gestures will be taken as inputs by the sensor, making it easier for them to understand.

This is a machine to human interaction system which includes Kinect sensor and Matlab for processing the data given as input.

There have been multitudinous researches done till date, but this paper provides a direct and flexible system for deaf and dumb people. It extracts voice from the human gesture of sign language as well as generates images and texts depending upon the input gestures given to the system. The very first step is to give the input as gesture data to the Kinect sensor, by this it senses the data and a 3-D image are created. This data is then transferred to Matlab where it is interfaced through the programming along with image processing and feature extraction using different segmentations and Hidden Markov Model (HMM) algorithm. From the complete segmented body, only the image of the hand is cropped, the gesture of that hand is then equated with the available image in the database and if they match the speech and text is obtained as output making it easier for the

common people to understand it. By this, the disabled people will be confident enough to express their views anywhere and everywhere despite physically challenged.

2. Literature Review

The Sign Language detection is considered an efficient way by which physically challenged people can communicate. Many researchers have studied and investigations are done on different algorithms to make the process easier.

Gunasekaran and Manikandan [1] have worked on a technique using PIC Microcontroller for detection of sign languages. The authors stated that their method is better as it solves the real time problems faced by the disabled ones. Their work involves extraction of voice as soon as any sign language is detected.

Kiratey Patil et al. [2] worked on detection of American Sign Language. Their work is based on accessing American Sign Language and converting into English and the output flashes on LCD. This way their work may omit the communication gap between common and disabled ones.

Tavari et al. [3] worked on recognition of Indian Sign languages by hand gestures. They proposed an idea of recognizing images formed by different hand movement gestures. They proposed an idea of recognizing images formed by different hand movement gestures. They used a web camera in their work. For identifying signs and translation of text to voice, Artificial Neural Network has been used.

Simon Lang [4] worked on Sign Language detection. He proposed a system that uses Kinect sensor instead of a web camera. Out of nine signs performed by many people, 97% detection rate have been seen for eight signs. The important body parts are sensed by Kinect sensor easily and Markov Model is used continuously for detection.

Sign Writing system proposed by Cayley et al. [5] deals with procuring in helping deaf people by using stylus and screen contraption for the written literacy in Sign Language. They have provided databases for enhancing the studies in another paper so that the sequence of the characters can be stored and retrieved in order to signify the sign language and then the editing could be done. In order to enhance their work on sensing algorithm, they are further researching on it.

According to Singha and Das [6], several Indian sign languages have been acknowledged by the process of skin filtering, hand cropping feature extraction and classification by making use of Eigen value weighted Euclidean distance. Hence out of 26 alphabets, only dynamic alphabets 'H & J' were not

taken into account & they will be considered in their future studies.

According to Xiujuan Chaivgftxtgf.,njoh et al. [7] for hard and body tracking 3-D motion by using Kinect Sensor is more effective and clear. This makes sign language detection easier.

According to our earlier work [8], the efficiency was 92% but now our efficiency has increased to 94.5%. We have used very simple algorithm here rather than using FCM.

From the above studies, it has been observed that few methods are only proposed for hand gestures recognition and few are only for feature extraction. Also from the above done survey, it is understood that no precise idea about feature extraction in easiest way is mentioned. But this problem is solved in our study. We have proposed an algorithm and used HMM technique also. Gestures are identified easily, that information is then matched with our preset databases and voice is extracted. This process enables the common people to understand the sign language easily.

3. Methodology

We have proposed an algorithm which follows the following steps:

1. After the detection of the body in front of the Microsoft X-box Kinect Sensor as shown in Fig.1, it locates the joints of the body by pointing it out and hence we get a skeletal image.



Fig. 1 Kinect Sensor

2. Then the segmented image of the body is formed from the skeletal image. The area of the hands where the signs are captured are cropped out of the whole segmented image. Then the cropped image is converted into dots and dashes. The length of the dash is 4 unit and the spacing between the two lines of dashes is also 4 units.
3. Through observation, we found that wherever the length of the dashes is greater than 4 units it resembles the image of the cropped hand.

4. In our proposed algorithm, we have taken the concept of loops, it is used to detect the black points that are the space between the dashes. This detection of black points determines the position of dashes by successive subtraction of points in the iterations which is going on and on.
5. The basic algorithm behind this work is that after successive subtraction of points if the value is equal to 4 then there is no data and if the value is greater than 4 then there is the actual image of the cropped hand.
6. Now, here arises a problem that how we detect the location of fingers. So the algorithm behind this is based on the formation of matrices of the black points detected earlier. In an iteration, the matrix coordinates of the finger are four times the number of lines of dashes.
7. To highlight the fingers, we plotted star point of the same coordinates as of the fingers and for more precise feature extraction the image processing is followed by filtration of the image that means the star points plotted on the figures go through following conditions.
 - If they fall on a straight line either horizontal or vertical.
 - If they fall on a constant slope.
 - If they fall within 4-unit coordinate difference.
8. Then it eliminates the identical points which we call as garbage point. Now we get the filtered image but the process of feature extraction continues for identification of fingers that whether it is an index finger, middle finger, ring finger, little finger or thumb.

Following is the Table 1 which shows the range of coordinates in which the fingers are detected.

Table 1: Range of coordinates of fingers

S No.	Finger	Range of coordinates
1	Index	80-88
2	Middle	60-68
3	Ring	44-52
4	Little	32-40
5	Thumb	104-112

The flow chart of the above algorithm is shown in following Fig.2:

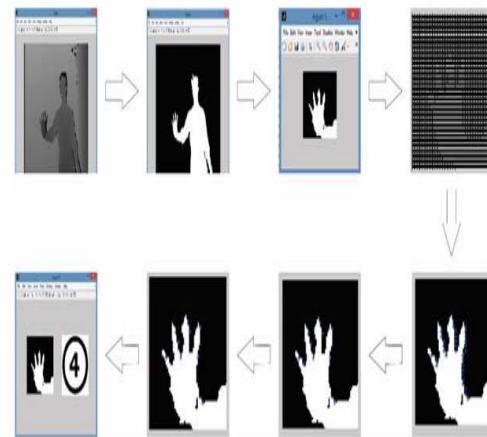


Fig. 2 Flow chart of the proposed algorithm

4. Hidden Markov Model

HMM [5], [9] is the algorithm which says that the actual stages of the work continued in the system is not visible the final output after the whole processing is only visible.

HMM works on probability and it uses a hidden variable of any input data and select them for various observations and then process all those variables through Markov process. HMM undergoes four stage process:

- Filtering: This state involves the computation which takes place during the hidden process of the given statistical parameters.
- Smoothing: This state does the same work as the filtering process but works in between the sequence wherever needed.
- Most likely Explanation: This state is different from the above two states. It is generally used whenever HMM is exposed to a different number of problems and to find overall maximum possible state sequences.
- Statistical Significance: This state of HMM is used to obtain statistical data and evaluate the data of the possible outcome.

5. Result

Finally, after the detection of the whole image of fingers or we can say that a complete hand ANDing operation continues in Matlab for the final output.

The detected image of the sign is searched in the database for its meaning and as and when the match is found search is complete and we get the final output along with the image of the meaning of sign and its voice.

We can take the example as, if number 4 is to be detected by the Kinect sensor then the person gestures 4 using his hands. The Kinect captures the skeletal image of the body as shown in the Fig. 3.

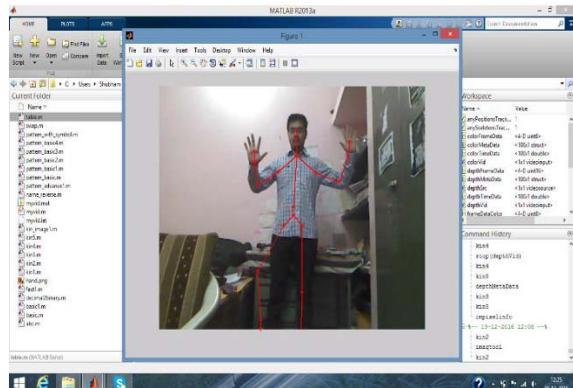


Fig. 3 Skeletal image

After skeletal image, the image is converted into depth image as in Fig. 4.

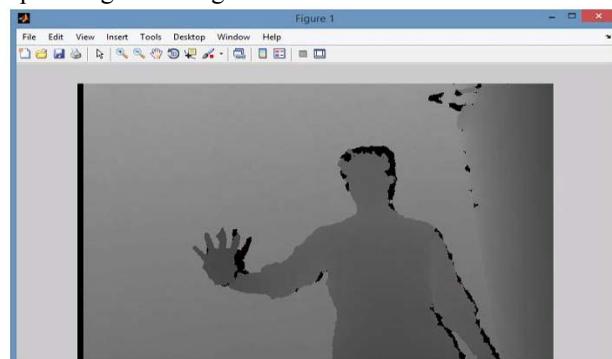


Fig. 4 Depth image

Then the image is converted into its segmented image as shown in Fig. 5.

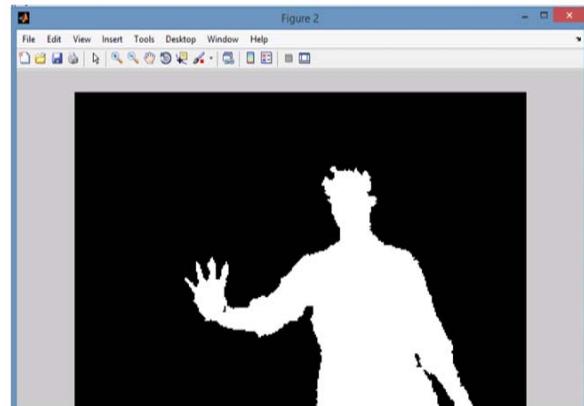


Fig. 5 Segmented image

The image of the hand is cropped from the segmented image as shown in Fig. 6.

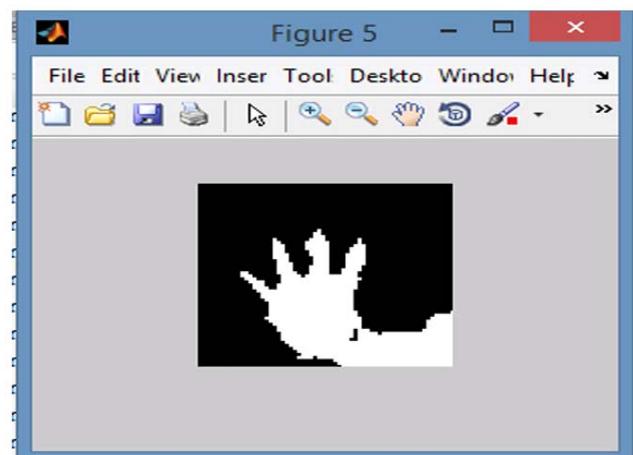


Fig. 6 Cropped image

The cropped image is then converted into a figure with dots and dashes as shown in Fig. 7.

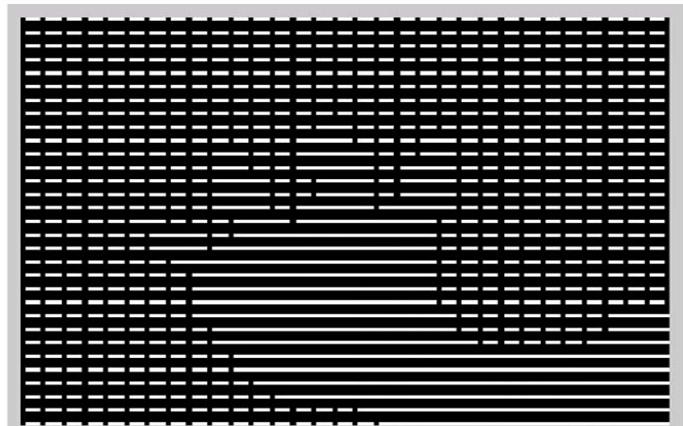


Fig. 7 Image with dots and dashes

The filtration of the figure after star marking it to detect the fingers is done in 3 steps as shown in Fig. 8.

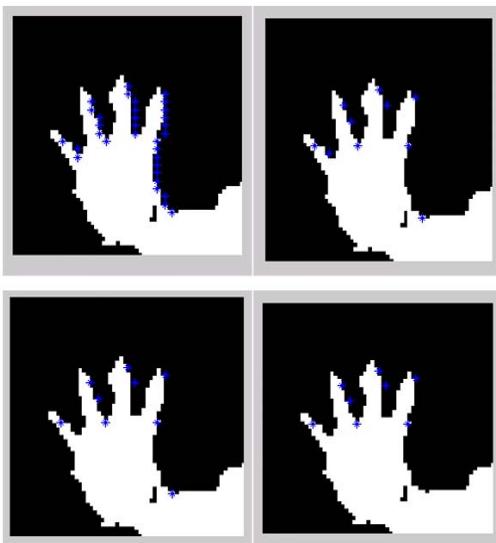


Fig. 8 Filtration of figure after star marking

The detailed information of the fingers detected are shown in command window which is shown in Fig. 9.

```
Command Window
fingure is between line 13 between column 11 and 12
fingure is between line 13 between column 16 and 17
fingure is between line 13 between column 20 and 21
fingure is between line 14 between column 11 and 12
fingure is between line 14 between column 16 and 17
fingure is between line 14 between column 20 and 21
fingure is between line 15 between column 11 and 12
fingure is between line 15 between column 16 and 17
fingure is between line 15 between column 20 and 21
fingure is between line 16 between column 6 and 7
fingure is between line 16 between column 12 and 13
fingure is between line 16 between column 19 and 20
fingure is between line 17 between column 8 and 9
fingure is between line 17 between column 19 and 20
fingure is between line 18 between column 8 and 9
fingure is between line 18 between column 19 and 20
fingure is between line 19 between column 19 and 20
fingure is between line 20 between column 19 and 20
fingure is between line 21 between column 19 and 20
fingure is between line 22 between column 19 and 20
fingure is between line 23 between column 20 and 21
fingure is between line 24 between column 20 and 21
fingure is between line 25 between column 21 and 22
little finger detected
ring finger detected
middle finger detected
index finger detected
% sign represent <>> |
```

Fig. 9 Detailed information of fingers detected

After the filtration is done the database is searched for its match and hence we get an output in the form of image as shown in Fig. 10(a) and voice which is plotted in the form of the histogram as shown in Fig. 10(b).



Fig. 10(a) Output in the form of image

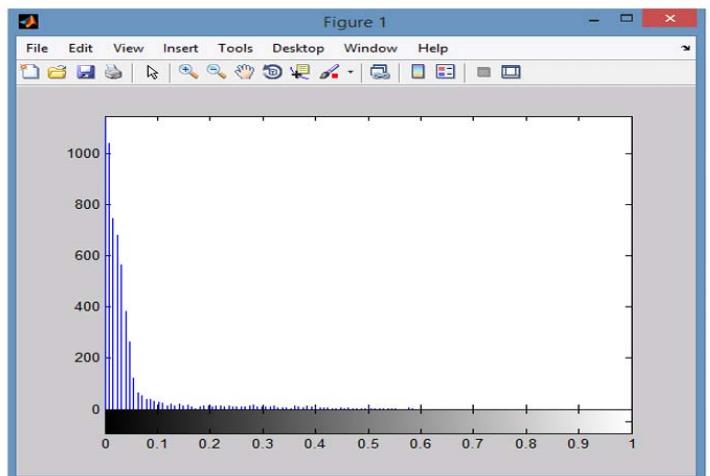


Fig. 10(b) Voice in the form of histogram

Overall output of various input given to the system shown in Fig.11.

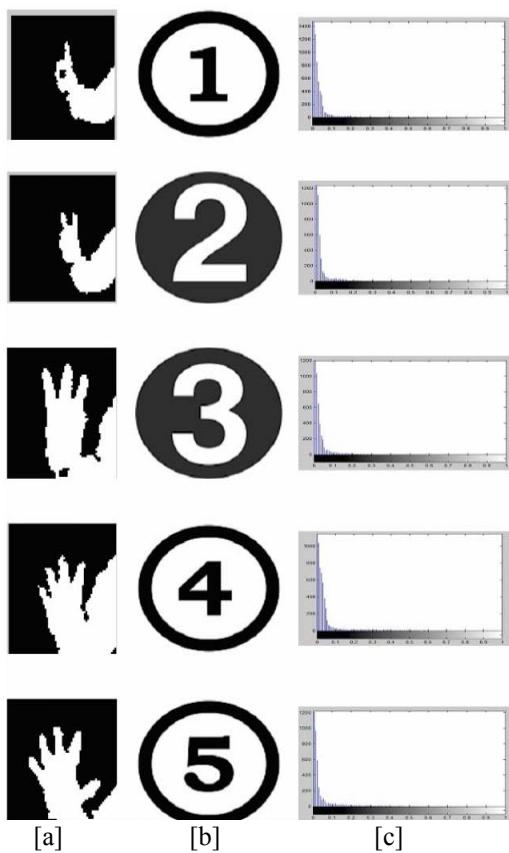


Fig. 11 [a] Cropped images, [b] Image of the meaning of signs, [c] Histogram image of voice output

Detailed analysis of the various outputs with respect to given inputs have been tabulated in Table 2.

Table 2: Detailed Analysis

S No.	No. of correct attempts	No. of wrong attempts	Accuracy(%)
1	50	0	100
2	50	0	100
3	49	1	96
4	48	2	92
5	49	1	96
6	49	1	96
7	47	3	88
8	47	3	88

Total no. of attempts = 50

Accuracy = (No. of correct attempts – No. of wrong attempts)/Total no. of attempts

Hence we get the total accuracy as 94.5%.

6. Conclusions and Future work

With references to all the earlier studies, our work provides the better accessibility with the simpler algorithm and more precise output.

Since programming is done for the detection of left hand the coordinates are taken accordingly. Our algorithm gives all the relevant information about the coordinates of each and every finger detected.

This system is very flexible and user-friendly as the user can be of any age, gender, size or color, the results will be same. But the intensity of light and distance of the body from the sensor affects the efficiency. For working effectively with the devices it is suggested to keep the Kinect sensor at a height of about 62 cm from the ground and the body to be detected should be distanced at about 90cm.

In our earlier work [8] the exact location of the fingers and the detailed information about them has not been determined, so we have overcome these problems here.

The star points that we have marked earlier is not completely filtered, very few points still remain there. Sometimes misinterpretation of detected fingers also occurs but its possibility is one out of ten.

Acknowledgments

We would like to extend our gratitude and sincere thanks to all people who have taken a great deal of interest in our work and helped us with a valuable suggestion.

References

- [1] K. Gunasekaran and R. Manikandan, "Sign language to speech translation system using pic microcontroller", International Journal of Engineering and Technology, vol. 05, no. 02, pp. 1024–1028, April 2013.
- [2] K. Patil, G. Pendharkar, and G. N. Gaikwad, "American sign language detection", International Journal of Scientific and Research Publications, vol. 04, pp. 01–06, November 2014.
- [3] N. V. Tavari, A. V. Deorankar, and P. N. Chatur, "Hand gesture recognition of Indian sign language to aid physically impaired people", International Journal of Engineering Research and Applications, pp. 60–66, April 2014.
- [4] S. Lang, "Sign language recognition with Kinect", Master's thesis, Freie University Berlin, 2011.

- [5] C.-Y. Kao and C.-S. Fahn, “A human-machine interaction technique: hand gesture recognition based on Hidden Markov Models with trajectory of hand motion”, Elsevier Advanced in Control Engineering and Information Science, vol. 15, pp. 3739–3743, 2011.
- [6] J. Singha and K. Das, “Indian sign language recognition using Eigen- value weighted Euclidian distance based classification technique”, International Journal of Advanced Computer Science and Applications, vol. 04, no. 02, pp. 188–195, 2013.
- [7] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, “Sign language recognition and translation with Kinect”, in IEEE International Conference on Automatic Face and Gesture Recognition, April 2013.
- [8] N.B.Bahadure, Sankalp Verma, Shubham Juneja, Chhaya Chandra, D.K.Mishra, and P.D. Mahapatra, “Sign Language Detection with Voice Extraction in Matlab using Kinect Sensor”, in International Journal of Computer Science and Information Security, vol.14, no. 12, pp.858-863, December 2016.
- [9] T. Starner and A. Pentland, “Visual recognition of American sign language using Hidden Markov models”, in Proceedings of International workshop face and gesture recognition, 1995, pp. 189–194.
- [10] S. R. Ganapathy, B. Aravind, B. Keerthana, and M. Sivagami, “Conversation of sign language to speech with human gestures”, in Elsevier 2nd International Symposium on Big Data and Cloud Computing, vol. 50, 2015, pp. 10–15.
- [11] M. Boulares and M. Jemni, “3d motion trajectory analysis approach to improve sign language 3d-based content recognition”, in Elsevier Proceedings of International Neural Network Society Winter Conference, vol. 13, 2012, pp. 133–143.
- [12] D. Vinson, R. L. Thompson, R. Skinner, and G. Vigliocco, “A faster path between meaning and form? Iconicity facilitates sign recognition and production in British sign language”, Elsevier Journal of Memory and Language, vol. 82, pp. 56–85, March 2015.
- [13] K. Cormier, A. Schembri, D. Winson, and E. Orfanidou, “A faster path between meaning and form? Iconicity facilitates sign recognition and production in British sign language”, Elsevier Journal of Cognition, vol. 124, pp. 50–65, May 2012.
- [14] H. Anupreethi and S. Vijayakumar, “Msp430 based sign language recognizer for dumb patient”, in Elsevier International Conference on Modeling Optimization and Computing, vol. 38, 2012, pp. 1374–1380.
- [15] C. Guimaraes, J. F. Guardez, and S. Fernandes, “Sign language writing acquisition- technology for writing system”, in IEEE Hawaii International Conference on System Science, 2014, pp. 120–129.
- [16] C. Guimaraes, J. F. Guardez, S. Fernandes, and L. E. Oliveira, “Deaf culture and sign language writing system- a database for a new approach to writing system recognition technology”, in IEEE Hawaii International Conference on System Science, 2014, pp. 3368–3377.
- [17] J. S. R. Jang, C. T. Sun, and E. Mizutani, Neuro - Fuzzy and Soft Computing. Eastern Economy Edition Prentice Hall of India, 2014.
- [18] S. C. Pandey and P. K. Misra, “Modified memory convergence with fuzzy PSO”, in Proceedings of the World Congress on Engineering, vol. 01, 2 - 4 July 2007.
- [19] M. S. Chafi, M.-R. Akbarzadeh-T, M. Moavenian, and M. Ziejewski, “Agent based soft computing approach for component fault detection and isolation of CNC x - axis drive system”, in ASME International Mechanical Engineering Congress and Exposition, Seattle, Washington, USA, November 2007, pp. 1–10.
- [20] S. C. Pandey and P. K. Misra, “Memory convergence and optimization with fuzzy PSO and ACS”, Journal of Computer Science, vol. 04, no. 02, pp. 139–147, February 2008.
- [21] N. B. Bahadure, A. K. Ray, and H. P. Thethi, “Performance analysis of image segmentation using watershed algorithm, fuzzy c – means of clustering algorithm and Simulink design”, in IEEE International Conference on Computing for Sustainable Global Development, New Delhi, India, March 2016, pp. 30–34.
- [22] H.-D. Yang, “Sign language recognition with Kinect sensor based on conditional random fields”, Sensors Journal, vol. 15, pp. 135–147, December 2014.

Computer Science Research Methodologies

Omankwu, Obinnaya Chinecherem, Nwagu, Chikezie Kenneth, and Inyama, Hycient

**¹ Computer Science Department, Michael Okpara University of Agriculture, Umudike
Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com**

**² Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
Nwaguchikeziekenneth@hotmail.com**

**³ Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka
Anambra State, Nigeria.**

ABSTRACT

This paper discusses the several research methodologies that can be used in Computer Science (CS) and Information Systems (IS). The research methods vary according to the science domain and project field. However a little of research methodologies can be reasonable for Computer Science and Information System.

Keywords- Computer Science (CS), Information Systems (IS), Research Methodologies.

1. INTRODUCTION

Science is an organized or systematic body of knowledge (Jain, 1997). Science embraces many different domains however those domains are related. Logic and mathematics sciences are the core of all sciences. From there and descending it emerge the natural sciences such as physic, chemistry and biology. In the next come the social sciences.

As (Denning, 2005) reported Computer Science (CS) should not be called as a science. Although CS is definitely a recent discipline, few will still argued that it is not provided with attributes to be qualified as a science. Computer Science has its specificity and has its bases in logic and mathematics.

However CS is transversal to very different domains in science. To understand which research methodologies that can be used in CS and IS we have to understand the differences between CS and IS.

Computer science (CS) characterized as an empirical discipline, in which each new program can be seen as an experiment, the structure and behavior of which can be studied (Allen ,2003). In particular, the field of CS is concerned with a number of different issues seen from a technological Perspective, e.g. theoretical aspects, such as numerical analysis, data Structures and algorithms; how to store and manipulate, the relationship between different pieces of software and techniques and tools for developing software .The field of Information Systems (IS) is concerned with the interaction between social and technological issues (Allen ,2003).

In other words, it is a field which focuses on the actual “link” between the human and social aspects (within an organization or other broader social setting), *and* the hardware, Software and data aspects of information technology (IT). In the next sections we will discuss the different methods of research methodology and its reasonability for the CS and IS domains.

Abstraction in Computer Science

Computer scientists are often thought to labor exclusively in a world of bits, logic circuits and microprocessors. Indeed, the foundational concepts of computer science are described in the language of binary arithmetic and logic gates, but it is a fascinating aspect of the discipline that the levels of abstraction that one can lay upon this foundational layer are limitless, and make possible to model familiar objects and processes of every day life entirely within a digital world. When digital models are sufficiently realistic, the environments they inhabit are called virtual worlds. So today, of course, there are virtual libraries, virtual shopping malls, virtual communities, and even virtual persons, like the digital version of actor Alan Alda created in an episode of PBS's *Scientific American Frontiers*.

Complex virtual worlds such as these are made possible by computer scientists' ability to distance themselves from the mundane and tedious level of bits and processors through tools of abstraction. To abstract is to describe something at a more general level than the level of detail seen from another point of view. For example, an architect may describe a house by specifying the height of the basement foundation, the location of load-bearing walls and partitions, the R-factor of the insulation, the size of the window and door rough openings, and so on. A realtor, however, may describe the same house as having a certain number of square feet, a certain number of bedrooms, whether the bathrooms are full or half, and so on. The realtor's description leaves out architectural detail but describes the same entity at a more general level, and so it is an abstraction of the architect's description. But abstraction is relative. For example, the architect's

description is itself an abstraction when compared to a metallurgist's description of the nails, screws, and other fasteners making up the house, and the botanist's description of the various cellular properties of the wood it contains.

The computer scientist's world is a world of nothing but abstractions. It would not be possible to create the complex virtual worlds described above if the only things computer scientists could talk about were bits, bytes, and microcircuits. One can give an accurate idea of what computer scientists do by describing the abstraction tools they use. Now to characterize computer science as involved with abstractions seems to claim for it a place alongside mathematics as a purely formal endeavor. But the general trends in all programming are toward higher-quality software by abstracting away from the lower-level concepts in computer science and toward the objects and information that make up the real world. This is a kind of abstraction that is fundamentally different from that which takes place in mathematics. Understanding the difference is crucial in avoiding the persistent misconception by some that computer science is just a branch of pure mathematics. Both mathematics and computer science are marked by the introduction of abstract objects into the realm of discourse, but they differ fundamentally in the nature of these objects. The difference has to do with the abstraction of form versus the abstraction of content.

Not all philosophers of mathematics agree with Carnap that mathematics has only linguistic utility for scientists, but there is agreement on the nature of mathematical abstraction being to remove the meanings of specific terms. M. Cohen and E. Nagel, for example, present a set of axioms for plane geometry; remove all references to points, lines, and planes; and replace them with symbols used merely as variables. They then proceed to demonstrate a number of theorems as consequences of these new axioms, showing that pure deduction in mathematics proceeds with terms that have no observational or sensory meaning. An axiom system may just happen to describe physical reality, but that is for experimentation in science to decide

Research Methods

Before we start in discuss the different types of research methodologies we have to define the research. In an academic context, research is used to refer to the activity of a diligent and systematic inquiry or investigation in an area, with the objective of discovering or revising facts, theories, applications etc. The goal is to discover and disseminate new knowledge. There are several methods that can be used in CS and IS in next subsection we will show these methodologies.

Experimental Method

Experimental shows the experiments that will occur in order extract results from real world implementations. Experiments can test the veracity of theories. This method within CS is used in several different fields like artificial neural networks, automating theorem proving, natural languages, analyzing performances and behaviors, etc. It is important to restate that all the experiments and results should be reproducible. Concerning, for example, network environments with several connection resources and users, the experiments are an important methodology Also in CS fields and especially IS

fields that take in consideration the Human- Computer Interaction. It is mandatory the usage of experimental approaches. If we use the experimental method in IS field we may need to use some methods or tools in conjunction with the experimental method. These methods or tools used to support and prove the legibility of the developed project. For example if a student wants to implement new social network with new concepts or develop an existing social network, who he can measure the legibility of his implementation? The answer of this question consists of two parts according to the nature of the project. The technical issue is the first part of the project that can be tested by benchmarks like (availability, reliability, scalability, stability, etc.). The usability of the project is the second part of testing that needs a feedback from the users of the system; the results for the second part can be obtained by the statistical analysis of a questionnaire which is a tool the used in conjunction with the experimental method.

Simulation Method

Simulation method used especially in CS because it offers the possibility to investigate systems or regimes that are outside of the experimental domain or the systems that is under invention or construction. Normally complex phenomena that cannot be implemented in laboratories evolution of the universe. Some domains that adopt computer simulation methodologies are sciences such as astronomy, physics or economics; other areas more specialized such as the study of non-linear systems, virtual reality or artificial life also exploit these methodologies. A lot of projects can use the simulation methods, like the study of a new developed network protocol. To test this protocol you have to build a huge network with a lot of expensive network tools, but this network can't be easily achieved. For this reason we can use the simulation method.

Theoretical Method

The theoretical approaches to CS are based on the classical methodology since they are related to logic and mathematics. Some ideas are the existence of conceptual and formal models (data models and algorithms). Since theoretical CS inherits its bases from logic and mathematics, some of the main techniques when dealing with problems are *iteration, recursion and induction*.

Theory is important to build methodologies, to develop logic and semantic models and to reason about the programs in order to prove their correctness. Theoretical CS is dedicated to the design and algorithm analysis in order to find solutions or better solutions (performance issues, for example). Encompassing all fields in CS, the theoretical methodologies also tries to define the limits of computation and the computational paradigm. In other words we can say that we can use the theoretical method to model a new system. However the theoretical method can help in finding new mathematical models or theories, but this method still needs other methods to prove the efficiency of the new models or theories. For example when a student need to develop a new classifier in AI by using the mathematical representation and theoretical method, he need to prove the efficiency of this model <https://using one of the previous methods>

Conclusion

In this paper we try to differentiate between the domains of science and CS and IS to understand the best methods that can be used in CS and IS. Each project in CS or IS have its free nature so the paper give examples of different kinds of projects in CS and IS and the proper research methodologies that can be used in these projects.

References

1. R. K. Jain and H. C. Triandis: Management of Research and Development Organizations: Managing the Unmanageable. John Wiley & Sons, (1997).
2. Gordana Dodig-Crnkovic: Scientific Methods in Computer Science. Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden (2002).
3. Denning Peter J.: Is Computer Science Science?. COMMUNICATIONS OF THE ACM, Vol. 48, No. 4 (2005).
4. Allen Newell, Herbert A. Simon: Computer Science as Empirical Inquiry: Symbols and Search. Communication. ACM 19(3): 113-126(1976).
5. Suprateek Sarker, Allen S. Lee: Using A Positivist Case Research Methodology To Test Three Competing Theories-In-Use Of Business Process Redesign. J. AIS (2001).

Recovery of RGB Image from Its Halftoned Version based on DWT

Tasnim Ahmed

Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
tasnim.cse20@gmail.com

Md. Habibur Rahman

Department of Computer Science & Engineering
Jahangirnagar University
Dhaka, Bangladesh
habib.cse.ju@gmail.com

Md. Imdadul Islam

Department of Computer Science & Engineering
Jahangirnagar University
Dhaka, Bangladesh
imdad@juniv.edu

Abstract— Halftoning of image is a way of compressing both RGB and grayscale image where instead of continuous levels or tone of pixels, only two discrete levels of pixels are considered. Actually a halftone image resembles a binary image in context of bits of pixels but the size and shape of pixels are modified to make it better in visualization. In this paper, we used two dimensional filtering techniques and discrete wavelet transform (DWT) with thresholding to recover an RGB image from its halftoned version. We compared the original and recovered image based on six largest eigen values, the SNR in dB and cross-correlation co-efficient of Red, Green and Blue components. The algorithm we used here shows 94% or above similarity between original and recovered image. This paper is actually the extended version of the previous paper of grayscale image.

Keywords- Signal to noise ratio, 2D filtering, standard deviation, eigen value, cross-correlation coefficient.

I. INTRODUCTION

A halftone image is actually a binary image of different formats. In a gray scale binary image each pixel is represented by a binary bit 1 or 0 against white or black; provided size of each pixel is equal. A halftone image is also made up of dots but size is not equal like binary image. In a colored half-tone image the dots are variable in sizes, shapes, colors. The halftone image takes the opportunity to represent dark areas with large dots while small dots are used to represent lighter areas. Halftoning is widely used in display devices like newspaper printers, laser printers, even some computer screens to reduce the size of image. The spectrum of an image mainly consists of low, medium and high frequency components. Human eye is highly sensitive to low frequency components and in halftone images the low frequency component is approximately same as the continuous tone image. The high-frequency component is not correlated with the low-frequency component of an image and does not convey vital information of an image as discussed in [1]. The human visual system approximately acts as a low-pass filter hence a half-tone image gives the illusion of continuous tone image from a distance. On the other hand,

storage capacity and transmission time of an image is an important issue satisfied by such image. The two basic operations of halftoning are: dithering and error diffusion discussed in [1-3] and theirs inverse operation is found in [4]. Among several inverse algorithm methods: digital filtering method is shown in [5] and error diffusion method is analyzed in [6]. The quality of halftone image is analyzed in [7] using the concept of amplitude modulation (AM) and frequency modulation (FM).

In this paper we use error diffused halftone (error is diffused to surrounding pixels) under Floyd-Steinberg mask to produce halftone dots of image. Actually halftone is a lossy compression like JPEG (Joint Photographic Expert Group), hence its recovery is also lossy but the proposed method provides a good impression. The computation complexity is less and the process time used here is also too small compared to the other existing models at the expense of quality of the image. Here we combined filtering and discrete wavelet transform with thresholding technique to recover the RGB image from its lossy halftone image. We use two parameters: cross-correlation co-efficient and six eigenvalues to measure the similarity of original and recovered image. The RED, GREEN and BLUE plates of RGB image is used as the matrices of real number like [8] for comparison. Finally recovered image is de-noised using Discrete Wavelet Transform (DWT) of [9]. Similar work was done by the third author of the paper in [10] only for gray scale image, but this paper gives the extension of [10] for RGB image which is actually three times more complex compared to it.

The rest of the paper is organized as: section II provides the algorithm of conversion of an RGB image into colored halftone image and its recovering techniques, section III deals with the results based on analysis of section II and section IV concludes entire analysis.

II. SYSTEM MODEL

The algorithm to convert an RGB image into colored halftone image and way of recovery of lossy image is given below:

- 1) Read the original RGB image.
- 2) Separate R, G, and B components of the image.
- 3) Convert each component of the image into halftone image using Floyd-Steignberg algorithm. Let us denote each halftone component as R_h , G_h and B_h .
- 4) Reconstruct the halftone RGB image combining R_h , G_h and B_h .
- 5) Display both the original and halftone image for comparison.
- 6) Smoothen each component of the halftone image making convolution with 2D filtering (motion, average, disk, Gaussian). Combine the filter matrix to form smooth RGB image.
- 7) Display the smooth image for comparison.
- 8) Apply DWT on the filtered image with hard threshold to remove noise grain of the filtered image.

III. RESULTS

We consider four test image (RGB images) shown in figure 1 to 4. Each figure composed of four components: original image, half-tone image, recovered image after convolution and de-noised image (applying DWT). First of all, we compare the original and recovered images in context of eigen values. The six largest eigen values (λ_i ; $i = 1, 2, 3, \dots, 6$) of original and recovered image are evaluated for R, G and B components using MATLAB 16. For the combination of Gaussian filter and DWT, we made the comparison in tabular form shown in Table (I-IV). From all the tables, we can see that error is found less than 7% and at a glance, the recovered images resemble to the original images.

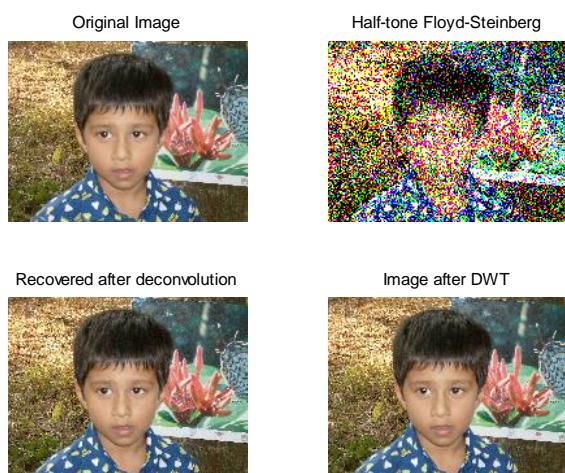


Figure 1. Siam as the test image

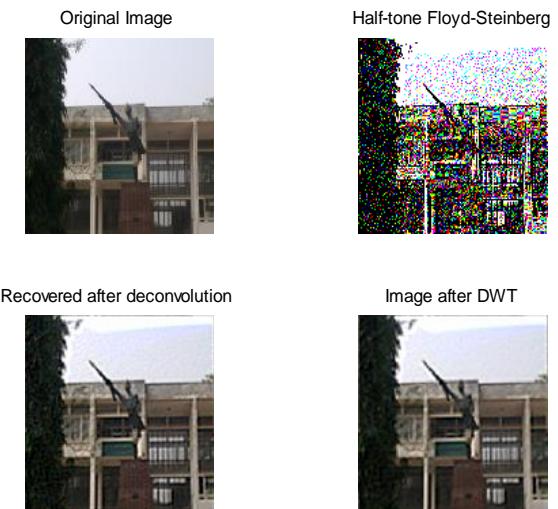


Figure 2. Statue of war '71 as the test image

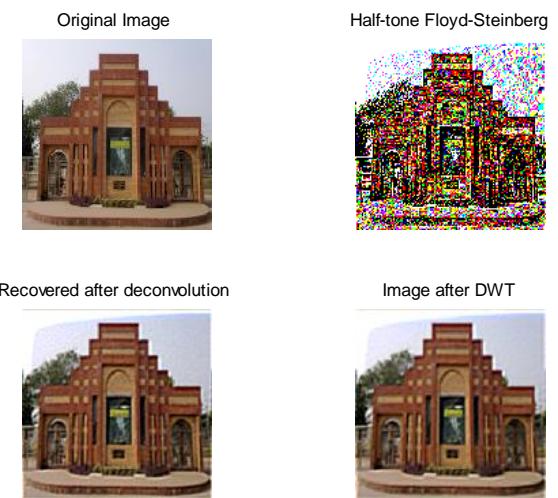


Figure 3. Jahangirnagar University gate as the test image



Figure 4. Vegetables as the test image

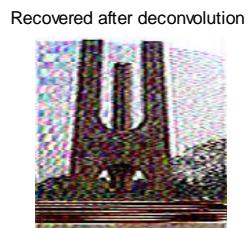
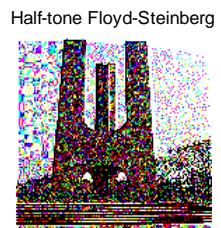
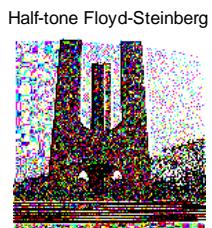


Figure 5. Motion Filter on the image of Victory Monument

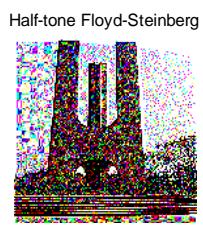


Figure 6. Gaussian Filter on the image of Victory Monument

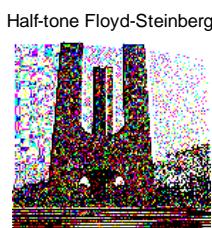


Figure 7. Disk Filter in Victory Monument Image

Figure 10. Gaussian Filter in Imdad Image

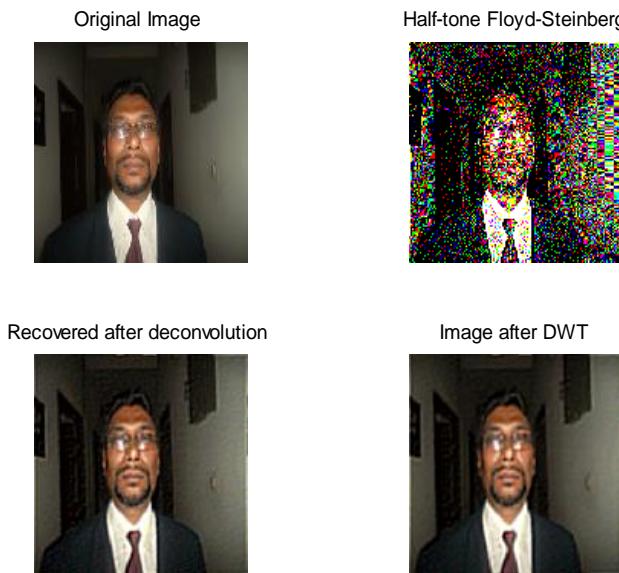


Figure 11. Disk Filter in Imdad Image

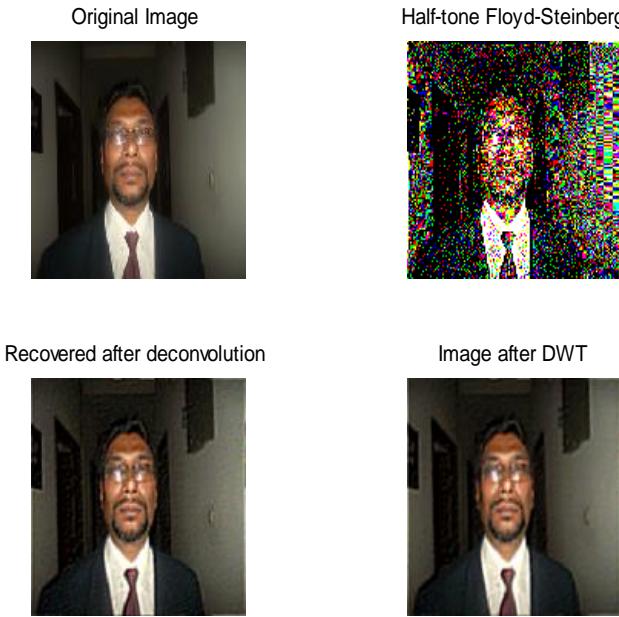


Figure 12. Average Filter in Imdad Image

In second stage, we compare the performance of four filters (used in convolution with half-tone image) of this paper, considering SNR in dB. In evaluating the SNR of recovered image, signal component is taken as the square sum of each pixel of the original image and noise is the mean square error between original recovered images. In this paper we made experiment on 100 images for comparison of four filters and only five are shown in Table V. From the Table V, the performance of the Gaussian, Disk and Average filters are very closed and better than Motion filter. Among the three, although performance is image dependent but Gaussian filter shows little

better results. The relative performance of four filters are shown in Figure 5 to 12.

TABLE I. COMPARISON OF IMAGES OF FIGURE 1

		λ_1	λ_2	λ_3
Original Image	Red	1.0000	0.2577	-0.0354 +0.0443i
	Green	1.0000	0.1773	-0.1105
	Blue	1.0000	-0.2095	0.1451
Recovered Image	Red	1.0000	0.2579	-0.0346 +0.0445i
	Green	1.0000	0.1775	-0.1109
	Blue	1.0000	-0.2103	0.1452
Error	Red	0	-0.0002	-0.0008-0.0002i
	Green	0	-0.0002	0.0004
	Blue	0	0.0008	-0.0001

λ_4	λ_5	λ_6
-0.0354 + 0.0443i	-0.0179 - 0.0434i	-0.0179 + 0.0434i
0.0205 + 0.0649i	0.0205 - 0.0649i	0.0169 + 0.0296i
0.0186 - 0.1010i	0.0186 + 0.1010i	0.0592
-0.0346 - 0.0445i	-0.0181 - 0.0431i	-0.0181 + 0.0431i
0.0211 + 0.0643i	0.0211 - 0.0643i	0.0164 - 0.0307i
0.0188 - 0.1008i	0.0188 + 0.1008i	0.0590
-0.0008+0.0888i	0.0002-.0003i	0.0002+.0003i
-0.0006+.0006i	-0.0006-.0006i	0.0005+.0603i
-0.0002-.0002i	-0.0002+.0002i	0.0002

TABLE II. COMPARISON OF IMAGES OF FIGURE 2

		λ_1	λ_2	λ_3
Original Image	Red	1.0	-0.0748	0.0660
	Green	1.0	-0.0820	0.0714
	Blue	1.0	0.0735	-0.0716
Recovered Image	Red	1.0	-0.0827	0.0802
	Green	1.0	-0.0896	0.0863
	Blue	1.0	0.0936	-0.0826
Error	Red	0	0.0079	-0.0142
	Green	0	0.0076	-0.01490
	Blue	0	-0.02010	0.0110

λ_4	λ_5	λ_6
0.0429 - 0.0379i	0.0429 + 0.0379i	0.0382
0.0355 - 0.0184i	0.0355 + 0.0184i	0.0386
0.0615 + 0.0000i	0.0355 + 0.0184i	0.0386
0.0448 - 0.0331i	0.0448 + 0.0331i	0.0395
0.0403 - 0.0167i	0.0403 + 0.0167i	0.0352
0.0567	0.0374 - 0.0059i	0.0374 + 0.006i
-0.001900-.0048i	-0.001900+.0048i	-0.0013
-0.004800-.0017i	-0.004800+.0017i	0.0034
0.004800	-0.001900+.0243i	0.0012-.006i

TABLE III. COMPARISON OF IMAGES OF FIGURE 3

		λ_1	λ_2	λ_3
Original Image	Red	1.0	-0.0363 - 0.0509i	-0.036 + 0.051i
	Green	1.0	-0.0545 - 0.0539i	-0.0545 + 0.052i
	Blue	1.0	-0.0711 - 0.0591i	-0.0711 + 0.059i
Recovered Image	Red	1.0	-0.0408 - 0.0518i	-0.0408 + 0.052i
	Green	1.0	-0.0585 - 0.0518i	-0.0585 + 0.052i
	Blue	1.0	-0.0792 - 0.0554i	-0.0792 + 0.055i
Error	Red	0	0.0045+0.0009i	0.0045-0.0009i
	Green	0	0.004-.0021i	0.0040+.0021i
	Blue	0	0.0081-0.0037i	0.0081+.0037i

λ_4	λ_5	λ_6
0.0423 + 0.0000i	0.0239 + 0.0343i	0.0239 - 0.0343i
0.0449 + 0.0537i	0.0449 - 0.0537i	0.0261 + 0.0128i
0.0312 - 0.0544i	0.0312 + 0.0544i	0.0398 - 0.0038i
0.0261 - 0.0406i	0.0261 + 0.0406i	0.0461 + 0.0000i
0.0461 - 0.0620i	0.0461 + 0.0620i	-0.0346 + 0.0000i
0.0392 - 0.0621i	0.0392 + 0.0621i	-0.0499 + 0.0000i
0.0162+0.0406i	-0.0022-.00630i	-0.022200-0.0343i
-0.0012+0.1157i	-0.0012-0.1157i	0.060700+0.0128i
-0.0080+0.0077i	-0.0080-0.0077i	0.089700-0.0038i

TABLE IV. COMPARISON OF IMAGES OF FIGURE 4

		λ_1	λ_2	λ_3
Original Image	Red	1.0	0.0015 - 0.0668i	0.0015 + 0.0668i
	Green	1.00	0.2378	-0.0628 +0.0982i
	Blue	1.0	0.2086	-0.1509 +0.0964i
Recovered Image	Red	1.0	0.0129 - 0.0707i	0.0129 + 0.0707i
	Green	1.0	0.2240	-0.0558 +0.1132i
	Blue	1.0	0.2066	-0.1307 +0.1078i
Error	Red	0	-0.011400+.0039i	-0.0114-0.0039i
	Green	0	0.0138	-0.0070-0.0150i
	Blue	0	0.0020	-0.0202-0.0114i

λ_4	λ_5	λ_6
-0.0486 + 0.0332i	-0.0486 - 0.0332i	0.0011 + 0.0480i
-0.0628 - 0.0982i	0.0161 + 0.0488i	0.0161 - 0.0488i
-0.1509 - 0.0964i	-0.0373 - 0.1084i	-0.0373 + 0.1084i
-0.0585 + 0.0364i	-0.0585 - 0.0364i	-0.0020 - 0.0582i
-0.0558 - 0.1132i	0.0024 - 0.0634i	0.0024 + 0.0634i
-0.1307 - 0.1078i	-0.0425 - 0.1353i	-0.0425 + 0.1353i
.009900-.0032i	0.0099+0.0032i	0.0031+0.1062i
-.007000+.0150i	0.0137+0.1122i	0.0137-.1122i
-0.020200+0.0114i	0.0052+0.0269i	0.0052-0.0269i

TABLE V. COMPARISON OF SNR OF FILTERS

Components	SNR of Motion Filter in dB	SNR of Gaussian filter in dB	SNR of Disk filter in dB	SNR of Average filter in dB	Images
Red	7.7747	16.4960	15.7634	16.0332	JU Gate
Green	6.5125	15.3026	14.5977	14.8784	
Blue	6.3847	15.2154	14.5184	14.7953	
Red	6.2119	14.8887	13.9019	14.3437	Siam
Green	5.9047	14.7530	13.8176	14.2172	
Blue	5.0557	13.9224	13.0362	13.4149	
Red	7.5025	17.2172	16.3457	16.6884	Victory Monument
Green	6.9845	16.7158	15.7945	16.1585	
Blue	7.4790	16.7468	15.8596	16.2185	

Red	8.6051	18.0540	17.4465	17.7843	Memorial of 1952
Green	8.4645	18.1133	17.4971	17.8392	
Blue	8.6390	18.1281	17.5007	17.8464	
Red	5.2553	15.5891	15.3160	15.5506	Imdad
Green	4.4307	15.3795	15.1600	15.3910	
Blue	3.5105	14.9206	14.7957	14.9966	

TABLE VI. COMPARISON OF CROSS-CORRELATION COEFFICIENT OF ORIGINAL AND RECOVERED IMAGES UNDER DIFFERENT FILTERS

Components	Motion ρ	Gaussian ρ	Disk ρ	Average ρ	Image
R	0.8730	0.9428	0.9364	0.9399	JU Gate
G	0.9062	0.9596	0.9544	0.9571	
B	0.9352	0.9749	0.9713	0.9731	
R	0.8551	0.9371	0.9284	0.9332	Siam
G	0.8092	0.9081	0.8982	0.9040	
B	0.8172	0.9146	0.9055	0.9107	
R	0.9137	0.9663	0.9603	0.9631	Victory Monument
G	0.9309	0.9737	0.9687	0.9711	
B	0.9427	0.9784	0.9744	0.9763	
R	0.9483	0.9843	0.9802	0.9820	Memorial of 1952
G	0.9462	0.9839	0.9796	0.9815	
B	0.9484	0.9846	0.9805	0.9823	
R	0.9262	0.9773	0.9732	0.9751	Imdad
G	0.9168	0.9760	0.9719	0.9739	
B	0.9095	0.9755	0.9712	0.9733	

In Table VI, we compared the cross-correlation coefficient, ρ of Red, Green and Blue components of original and recovered image. We use the same images and filters of previous table and get the performance like before.

IV. CONCLUSIONS

In this paper, we recovered RGB image from its halftoned version using combination of image filtering and DWT. Here, we worked on halftone under Floyd-Steinberg algorithm. Still we have scope to work on other halftone algorithms for comparison. Our analysis will be helpful to save image storage and to save transmission time of image where lossy image compression is applicable. The compression ratio of our technique is much higher than JPEG since each pixel of colored halftone image requires only 3 bits instead of 24 bits of RGB image. Next we can introduce mask block such a way that zeros are filled at high frequency components. Now applying convolution on each block of the image with the mask can further reduce the size of image since we can apply run length code on the image with huge zeros.

REFERENCES

- [1] Soren Hein and Avideh Zakhor, ‘Halftone to Continuous-Tone Conversion of Error-Diffusion Coded Images,’ IEEE Transactions on Image Processing, vol.4, no. 2, pp.208-216, February 1995
- [2] H.B. Kekre, Tanuja K. Sarode, Sanjay R. Sange, Pallavi Halarnkar, ‘New Half tone Operators for High Data Compression in Video-

- Conferencing,’ 2012 International Conference on Software and Computer Applications (ICSCA 2012), vol.41, pp.211-217, Singapore, 2012
- [3] Seong Jun Park, Mark Q. Shaw, George Kerby, Terry Nelson, Di-Yuan Tzeng, Kurt R. Bengtson, and Jan P. Allebach, ‘Halftone Blending Between Smooth and Detail Screens to Improve Print Quality With Electro photographic Printers,’ IEEE Transactions On Image Processing, vol. 25, no. 2, pp.601-614, February 2016
- [4] Kuo-Ming Hung, Ching-Tang Hsieh2, Cheng-Hsiang Yeh and Li-Ming Chen, ‘Watermarking-Based Image Inpainting Using Halftoning Technique,’ Journal of Applied Science and Engineering, vol.15, no. 1, pp. 79-88 (2012)
- [5] Z. Fan, ‘Retrieval of images from digital halftones,’ in Proceedings of the IEEE International Symposium on CircuitsSystems, pp. 313–316, 1992.
- [6] T. D. Kite, N. Damera-Venkata, B. L. Evans, and A. C. Bovik, ‘A fast, high-quality inverse halftoning algorithm for error diffused halftones,’ IEEE Transactions on Image Processing, vol.9, no.9, pp. 1583–1592, 2000
- [7] Ivan Pićjer, Dragoljub Novaković1, Uroš Nedeljković1, Nemanja Kašiković1, Gojko Vladić1, ‘Impact of Reproduction Size and Halftoning Method on Print Quality Perception,’ Acta Polytechnica Hungarica vol. 13, no. 3, pp. 81-100, 2016
- [8] Rafael G. Gonzalez, Richard E. Woods and Steven L. Eddins, ‘Digital Image Processing using MAT-LAB’, Pearson Education, Inc., 1st edition, Delhi, 2004
- [9] M. R. Banham and A. K. Katsaggelos, ‘Spatially Adaptive Wavelet-Based Multiscale Image Restoration,’ IEEE Trans. Image Processing, vol.5, no.4, pp. 619-634, April 1996
- [10] Hafsa Moontari Ali, Rokhsana Khanom, Sarnali Basak and Md. Imdadul Islam, ‘Recovery of GrayScale Image from Its Halftoned Version Using Smooth Window Function’, Jahangirnagar University Journal of Electronics and Computer Science, vol. 15, pp.15-22, June 2014

AUTHORS PROFILE



Tasnim Ahmed received her B.Sc. (Honors) and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh in 2016 and 2018, respectively. Currently, she is working as a faculty member in the Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Her research interest is focused on Image Processing, Convolutional Neural Networks, Medical Imaging.



Md. Habibur Rahman received his B.Sc. (Honors) and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh in 2016 and 2018 respectively. His research interest is focused on Image Processing and Machine Learning.



Md. Imdadul Islam has completed his B.Sc. and M.Sc Engineering in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh in 1993 and 1998 respectively and has completed his Ph.D degree from the Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh in the field of network traffic in 2010. He is now working as a Professor at the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. Previously, he worked as an Assistant Engineer in Sheba Telecom (Pvt.) LTD, a joint venture company between Bangladesh and Malaysia from September 1994 to July 1996. Imdadul Islam has good field experience in installation Radio Base Station and configuration of Switching Centers for both mobile and WLL. His research field is network traffic, wireless communications, cognitive radio, LTE network, wavelet transform, OFDMA, adaptive filter theory, ANFIS and array antenna systems. He has more than hundred and seventy research papers in national and international journals and conference proceedings.

Data Redundancy on Diskless Client using Linux Platform

B. S. Sonawane
Research Fellow,
Dept. of CSIT,
Dr. B. A. M. University,
Aurangabad
bssonawane@gmail.com

R. R. Deshmukh
Professor,
Dept. of CSIT,
Dr. B. A. M. University,
Aurangabad
rrdeshmukh.csit@bamu.ac.in

S. D. Waghmare
Research Fellow,
Dept. of CSIT,
Dr. B. A. M. University,
Aurangabad
waghmare.swapnil21@gmail.com

Pushpendra Chavan
Principal, Tech Support
Engineer, Red Hat India
Pvt. Ltd, Pune,
chavanpushpendra@gmail.com

ABSTRACT

This paper addresses the redundant array of inexpensive/independent disks (RAID) in the field of diskless clients' where the centralized Disk Less NFS Server is present to share the OS bit to diskless clients over TCP IP over Local area network. Disk less client technology is very much practical and useful where a cost efficient and low end clients can be made useful without presence of a local disk itself. The clients which has the whole low end Computer system i.e. Keyboard, mouse, CPU + motherboard, monitor and may or may not have a disk can still use the advantages of RAID system through the diskless client server in such a way that any disk faults can be tolerated online. The underlying present disk (if any) in diskless client can also be used for specific purposes.

General Terms

The Disk Less Client is a technology used for ease of administration. The scope of this paper will be only for Software oriented Virtual Disk Less clients.

Keywords

RAID, disk less clients, redundancy, Mirroring, Fault Tolerance, latency, hardware and software data storage, data syncing, master-slave disks.

INTRODUCTION

Redundant array of inexpensive/independent disks i.e. in simple words, RAID, is very much aware term in computing world where data is more important instead of cost.

The high value data must be kept alive without any corruption, so the only possible ways are to either get data backed up on regular basis, and confirm the backed up data is restorable or not, or get the RAID system in place.

RAID is nothing but the methodology to keep the data copy on another disks in order to tolerate the faults. Fault tolerance ratio depends upon the number of disks and architecture of RAID system in use. Centralized Server is running on top of the highly effective hardware, there won't be performance degradation and the Guest Image running on top of the server can be easily managed. This makes System Administrators identify/create/modify/deploy anything in Single Clicks. One doesn't need to worry about the HDD corruption of the Client machines because there won't be any HDD on the same. Single Instance hosted at Centralized server makes it simple to maintain secure and robust as well. As it will be cost effective, the Corporate Industries can invest more on Human

Resources thus better paying jobs. Understanding RAID and Disk less Clients in short:

RAID :

Redundant array of Inexpensive/Independent disks is nothing but an architecture of disks attached together to perform fault tolerance technique for data. RAID architectures can be managed by direct software or by hardware. Software Managed RAID Architectures are called as Software RAID (mainly maintained by Operating system and is an overhead for Operating system to manage it) whereas Hardware Controller Managed RAID Architectures are called as Hardware RAID, where a special device called as Hardware RAID Controller is placed on Motherboard situated in between the Data Bus and the Disks. This Hardware RAID controller is responsible for all data writes and data reads, so Operating system is free to invest it's resources somewhere else while being in execution.

1. LITERATURE REVIEW

Disk less clients are the hardware low end systems which can run a specified Operating system presented by Disk less client server system over the IP NFS protocol. The Disk less client server hosts the chrooted (OS inside OS) image of stripped down or required package set Operating system over the network using network file system protocol and a network dracut image. These clients will boot from Network Interface cards to get the IP information from one of the DHCP server, making them a part of the desired network. The same DHCP server along with IP information, will provide the TFTPBOOT path for the clients to boot the system with the pre-existing OS on Diskless client server network. The Complete Image of chrooted and exported Operating system will be copied to the Physical memory (RAM) of the clients and will be loaded. The login information, the file system and the data to be operated will be present on the server's chroot OS and clients can make use of it.

This paper addresses the Technical Collaboration of RAID Technologies and Disk Less Client Configuration altogether to form a fully functioning production ready setup.

Diskless Technology is the client computers will not use hard disk drive. These clients are connected with the high speed network, run the file system from the server's hard disk simulation. With the help to this system client can manage the resources easily from client side, but due piracy problem some resources cannot be directly access by the client. The virtualization technology offers applications an abstract view through interfaces of the hardware platform and resources. Virtualization has several benefits for enabling cloud

computing. The hypervisor is a program of virtualization which handles the protection among virtual machines; hence applications can be easily migrated on different virtual machines without isolation of host integrity. Virtual machine architecture having six major types i.e. Full Virtualization, Hardware assisted virtualization, Para Virtualization, Operating System Level Virtualization, Application Level Virtualization and Network virtualization, in this research paper we are using the concept of Network virtualization to implement the diskless client and zero hardware environment for computer laboratories[1]. Distributed resource manager concept helps us to improve the scalability of operating systems which will be access by the client during run time environment. The Operating system will be residing on the Host OS on Remote normal/RAID storage as per the requirements. This will be shared using NFS protocol and will make sure to be accessible outside the HOST Operating System [2]. The Network used here will be of 10GBPS at least, because the OS will be transferred over the network itself. This will be a local area network which will take care of the total transfer of the data over Ethernet. The Diskless Clients will boot from network and will request for the DHCP IP address by sending an ARP (Address Resolution Protocol) packets broadcast over the network. This DHCP server will provide the IP address using DORA process and along with the DHCP IP address, the next-server IP address will also be shared here which will be the IP of the same HOST OS which shares the Guest OS image using NFS [3] in this article operating system having dynamic IP address so that user not having rights to change the IP, after restarting the operating system IP can be changed every time.

The TFTP Server will host the PXE kernel as well as the Guest OS kernel along with the network initrd.img files in the menu listing. The PXE kernel will be transferred via tftp server and that kernel image will read out the menu file thus moving to the NFS shared image of the Guest OS. The NFS protocol will be used to share the Guest Image of the OS for Disk Less Clients. Once the menu listing is loaded, the system will transfer the NFS hosted Guest OS image over network. The Client machines are the machines which just have Central Processing Unit, Monitor, Keyboard, Mouse but not the HDD. The system will boot using network, so network boot enabled NICs are required too. In order to keep the Host OS secure and robust, the Proxy and firewall is required. The firewall will have ports open for DHCP, NFS and TFTP services [3].

Server virtualization brings revolutionary changes to the data centers and it will provide the high speed connectivity during accession of operating systems from remote locations, currently Amazon provides full operating systems as pay as peruse concept but not affordable to the small scale and other offices. Integrating servers by virtualization can significantly reduce the space and power consumption, enhances the IT services, simplify the hardware management so that virtualization term is very much popular in the business market.[4]. Cloud computing, and particularly the use of public clouds, brings advantages on the technical, environmental and business sides, allowing multiple under-utilized systems to be consolidated within fewer physical servers hosting them. A cloud provider can manage physical resources in a very efficient way by scaling on the several hundreds and thousands of customers [1, 4].

2. PROPOSED SYSTEM

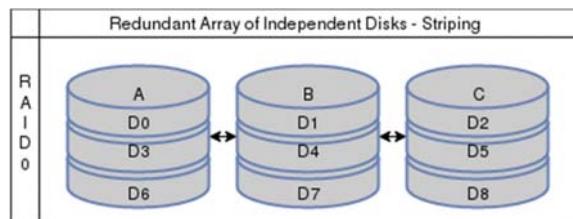


Figure 1. Proposed Diskless Systems Using RAID 0 Architecture.

RAID Architectures

The RAID configuration architecture depends on the way the disks are attached together and data is written on top of them. The major primary RAID architectures are as follows. Rest all RAID architectures are combination of following three major RAID architectures.

RAID 0 : Striping of Data.

- a. The RAID0, also called as Striping of Data RAID, architecture consists of minimum two or more disks.
- b. The motive here isn't for fault tolerance but to make Operating System think of many small disks as one through RAID0 architecture. I.e. maximum use of available disk space.
- c. Fault tolerance isn't present in RAID0
- d. The data is written across disks.
- e. Write and read operations both are faster.
- f. The architecture looks as follows.

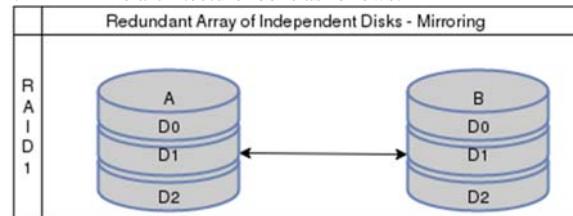


Figure 2. . Proposed Diskless Systems Using RAID 01 Architecture.

RAID 1 : Mirroring of Data

- a. The RAID1, also called as mirroring of Data, architecture consists of two disks by maximum.
- b. The motive here isn't the maximum disk space usage but to have maximum fault tolerance.
- c. Fault tolerance is 100%
- d. The data is copied across the other disk as a mirror copy.
- e. Write operation is slower than read operation.
- f. The architecture looks as follows.

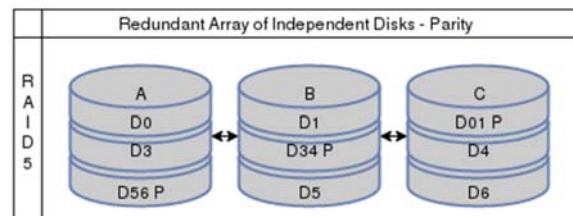
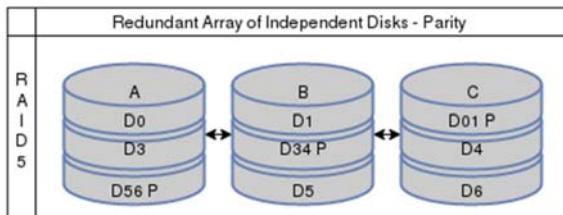


Figure 3. . Proposed Diskless Systems Using RAID 5 Architecture.

RAID 5 : Distributed Parity of Data

- a. The RAID5, also called as distributed Parity of Data, architecture consists of exact three disks .

- b. The motive here is to use the maximum disks with fault tolerance.
- c. Fault tolerance is 66%
- d. The data is wrote across two disks and an equivalent parity is written on third. This parity is written across each disk on sequential basis, so called as distributed parity.
- e. Write and read both are faster.
- f. The architecture looks as follows.

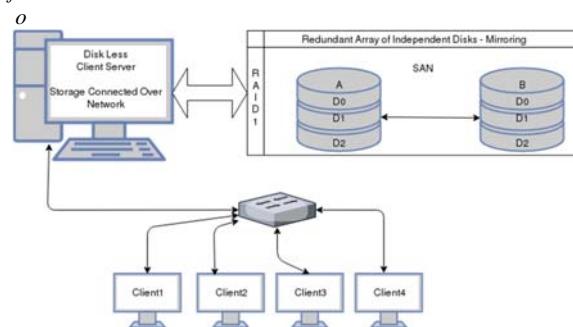


3. IMPLEMENTATION

In Order to have a working prototype of a Disk Less Client in-house, following requirements needs to be full filled. A

Diskless Client and RAID 1 :

Diskless client server's TFTPBOOT directory which hosts the chrooted Operating system image for booting purpose of the diskless clients, can be created as a RAID 1 - Mirroring Device which will give maximum fault tolerance over the site in case of any disk goes faulty. No downtime required for any operations here for recovery. The architecture will look as follows.



Disk Less Client Server Chrooted OS on a Storage Area Network connected through high fibre channel connection and in RAID1 System enables the system highly efficient for Fault tolerance making it one of the best design for maximum read speed over the clients. In above scenario, the OS itself is managing the RAID1 over the network for writing and exporting the OS.

A diskless Client machine can be created using Kickstart snippet below for fully functioning RAID 1 mounted under /var/lib/tftpboot directory as follows. Along with OS partitioning, the disks /dev/sdb and /dev/sdc is used for RAID1 to be mounted under /var/lib/tftpboot/ directory for diskless client chroot OS.

Kickstart Snippet for Automated Installation for Diskless Client's RAID volume

```
# Install bootloader on MBR of sda disk i.e. local disk of the
system.
bootloader --location=mbr --driveorder=sda,sdb,sdc --
append="crashkernel=auto rhgb quiet"
```

CentOS6 as a base machine is required to be used as a HOST OS which will serve as a Parent holding almost all the services which will be required by the Disk Less Client as shown in above Diagram. The same server will host the Local Repository of the Packages which will be deployed during GUEST image creation. The Services like DHCP (Dynamic Host Control Protocol), PXE (Pre-boot), TFTP and NFS will be running on the same HOST OS Machine too.

The Proof of Concept can be implemented as follows.

1. Install CentOS 6 machine with minimum 20GB free space and make sure it has a static IP address 192.168.0.254 (or whatever), we took 192.168.0.254/24 as the IP address for this Server
2. After installation, attach the CentOS 6 DVD to the machine, and mount it on the server.

```
# clear all disks i.e. sda which is local disk and sdb, and sdc
which are SAN block devices.
```

```
clearpart --drives=sda,sdb,sdc --initlabel
```

```
# create /boot, swap and pv.00 on /dev/sda as primary devices
for System Volume.
```

```
part /boot --size=1000 --ondisk=sda --asprimary --
fstype=ext4
part swap --size=1000 --ondisk=sda --asprimary --
fstype=swap
part pv.00 --size=10000 --ondisk=sda --grow --asprimary
```

```
# create a SystemVol volume group
volgroup SystemVol pv.00
```

```
# create three logical volumes for root, home and var
logvol / --vgname=SystemVol --size=6000 --name=rootvol
logvol /home --vgname=SystemVol --size=2000 --
name=homevol
logvol /var --vgname=SystemVol --size=5000 --name=varvol
```

```
# Another part of kickstart to create RAID1 on top of sdb and
sdc directly
part raid.0001 --size=5000 --grow --ondisk=sdb
part raid.0002 --size=5000 --grow --ondisk=sdc
raid pv.01 --device md1 --level=RAID1 raid.0001 raid.0002
```

```
# create a volume group out of the RAID volume pv.01 for
diskless client
volgroup DISKLESSVol pv.01
```

```
# Create and mount the RAID LVM on /var/lib/tftpboot for
100% fault tolerance.
logvol /var/lib/tftpboot/ --vgname=DISKLESSVol --size=6000
--grow --name=disklessclientvol
```

4.1 COMPARATIVE ANALYSIS

Table: 1 Comparative Analysis

Characteristics	Thin Client	Thick Client	Diskless Client using RAID
Low Watt Power Supply	Good	Poor	Good
Energy Efficiency	Good	Poor	Good
CPU Capacity	Poor	Good	Good
Cost	Medium	Poor	Excellent
Availability	Medium	Good	Good
Support	Good	Good	Good

We have checked the performance of diskless client using open source architecture along with existing technologies available in the market, we can connect more than 1000 systems where as other technologies having some limitation while connecting more than 100 systems, we have calculated the performance of the technologies such as thin client, thick client and diskless client supported with open source architecture as per the characteristics shown in the table 1.1

Many of the aspect related with the speed, Availability, Graphics, Energy our solution is good. The most important part to understand in the *POC* is the *dracut* network image. If the root partition is on a network drive, one has to have the network *dracut* modules installed to create a network aware *init image*. This *initramfs.img* gets downloaded using *tftp protocol* and then creates an environment in Physical Memory in order to create a feasible environment for mounting the further file systems.

Why we are promoting open source because we all know paid operating systems cost is not affordable to any section like school, colleges, universities, offices, industries etc. and this cost is increasing day by day, we need have a stable operating system and stable hardware cost also, but the scenario is different the cost of hardware is also not affordable, if we installed this solution to above said area it will save cost energy and affordable to the mass education and every on can

dream for free operating systems and every one can happy to learn the computer system without any trouble.

4. CONCLUSION

The design of Diskless client with the RAID 1 Mirroring will be highly recommended to use for its 100% fault tolerance and faster read operations. This will enable the users to get access to their data with highest speed and system will make sure that even if one disk gets corrupted, the data is still there for further operations.

5. REFERENCES

- [1] Marisol Garcia-Valls, Tommaso Cucinotta, Chenyang Lu, 2014. Challenges in real-time virtualization and predictable cloud computing, Journal of System Architecture 60 726-740. @2014 Elsevier B. V. All rights reserved.
- [2] Kulthida phapikhon, suchart khummanee, panida songram, chatklaw jareanpon, 2012. Performance Comparison of the Diskless Technology, 10 th Internaional Joint Conference on Computer Science and Software Engineering (JCSSE).
- [3] Shingo Takada, Akira Sato, Yasushi Shinjo, Hisashi Nakai, Akiyoshi Sugiki and kozo Itano,” @2013 IEEE. A P2P Approach to Scalable Network-Booting, Third International Conference on Networking and Computing.
- [4] Noki Tanida, Kei Hiraki, Mary Inaba, “Efficient disk-to-disk copy through long-distance high-speed networks with background traffic”, Fusion Engineering and Design, www.elsevier.com/locate/fusengdes.
- [5] Yaoxue zhang, Yuezhi Zhou, 2011. “separating and computation and storage with storage virtualization “ computercommunication, www.elsevier.com/locate/comcom
- [6] G. Clarco, M. Casoni, 2012. “On the Effectiveness of Linux Containers for network virtualization. Simulation modeling practice and theory. www.elsvier.com/locate/simpat.
- [7] Jinqian Liang, Xiaohong Guan, 2006 “A Virtual Disk Enviorment for providing file system recover” Science Direct. www.elsevier.com/locate/cose.

Enhanced Feature Analysis Framework for Comparative Analysis & Evaluation of Agent Oriented Methodologies

Omankwu, Obinnaya Chinecherem, Nwagu, Chikezie Kenneth, and Inyama, Hycient

**¹ Computer Science Department, Michael Okpara University of Agriculture, Umudike
Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com**

**² Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
Nwaguchikeziekenneth@hotmail.com**

**³ Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka
Anambra State, Nigeria.**

ABSTRACT

The objective of this paper is to provide an insight preview into various agent oriented methodologies by using an enhanced comparison framework based on criteria like process related criteria, steps and techniques related criteria, steps and usability criteria, model related or “concepts” related criteria, comparison regarding model related criteria and comparison regarding supportive related criteria. The result also constitutes inputs collected from the users of the agent oriented methodologies through a questionnaire based survey.

Keywords— Agents, Agent Oriented Methodology, Feature Analysis Framework, GAIA, PROMETHEUS, MESSAGE,

1. INTRODUCTION

The objective of this paper is to provide an insight preview into existing agent- oriented methodologies (AOM). Various agent oriented methodologies like GAIA, TROPOS, MAS-COMMONAKADS, PROMETHEUS, PASSI, ADELFE, MASE, RAP, MESSAGE and INGENIAS etc are available and are widely discussed. A comparison of five major agent oriented methodologies: GAIA, TROPS, PROMETHEUS, MESSAGE and MASE are presented in this paper. There had been various types of comparisons [1] done previously also by many researchers and software engineers, these comparisons are based upon certain different criteria [2] like process related criteria, steps and techniques related criteria, steps and usability criteria, model related or “concepts” related criteria, comparison regarding model related criteria and comparison regarding supportive related criteria. All these different comparisons cover almost all features of these methodologies like Application development life cycle support, coverage of life cycle, development approach, type of application domain, agent nature, ease of understanding of development steps etc. Ironically, the “best” methodology cannot be judged as these methodologies are application oriented and none of them can be considered as a perfect template or generalized framework for all kind of agent based applications. The careful evaluation of these methodologies can help developers in choosing the best methodology as per their application requirement.

THE COMPARISON FRAMEWORK

We adopted the feature analysis framework proposed by Tran, Low and Williams [3] as the basis as shown in the figure 1. The feature analysis framework constitutes four criteria: Model related criteria, Technique related criteria, Process related criteria and Supportive Features related criteria.

This framework is capable of assessing AOM (Agent Oriented methodologies) from both the dimensions of conventional system development methodologies and specific to AOSE. Also this framework is also capable of assessing the AOM at a multi-stage level. We are not using the full feature analysis as such but a modified version of the same has been used.

THE EXTENDED COMPARISON FRAMEWORK

The motive of extending the framework is to constitute the classical generic software engineering features [4] [5] [6] in addition to the elements specific to AOSE [7] [8] [9]. Moreover few features of object oriented software engineering are also compared in the framework. In our framework, we are using combination of major attributes of the all the criteria available under the feature analysis framework with the objective to compare the available features in the agent methodologies. These criteria constitute the respective attributes and features along with their description. It will help us assessing the methodologies on some specific guidelines. The details are as under.

A. *Model Related Criteria:* The model related criteria examine the capabilities and characteristics of the methodology’s models and notational components [10]. It also constitutes the concepts represented by the model [11] which is also a basis of our comparison framework. The concept property is divided into three further sub-sections: Internal properties, social properties and technical properties. It constitutes (a) AUTONOMY, which states that Agents can execute, operate and can be self-decisive of their own without any direct/external human intervention. Agents must have an inherent control on their internal state

which is dynamic in nature and can be modified by taking inputs from other agents in the environment. (b) REACTIVITY, which states that agents should respond in a consistent way towards changes occurring in the environment. The changes are triggered by the other agents present in the environment. (c) CONCURRENCY, which states that agents must interact with other agents simultaneously to achieve more than one goal. (d) PRO-ACTIVENESS, which states that agents must keep track of their goals evolving over time. Goals can evolve due to the changes in the environment. (e) ENVIRONMENT BELIEF, which states that agents must receive inputs from the environment, act accordingly and then may provide output to the environment, which can be used by other agents working in the environment. (f) COOPERATIVE BEHAVIOR, which states that Agents can request, respond, deny and even negotiate with other agents [12] in order to perform their individual goals and the system goals. (g) COMMUNICATION ABILITY, which states that agents can communicate directly, transitively, single directional (one to one) or multi-directional like a broadcast system [13]. (h) ACP (Agent Communication Protocol), which state that different agents communicate with each other by the means of message passing [12] [13]. These messages may be two fold also. A valid sequence of messages is required in order to achieve the goal(s). (i) ACL (Agent Communication Language), ACL provides agents with a mean of exchanging information and knowledge between them [13]. Using ACL, agents transport messages over the network using low level and high level protocols. (j) COMPLETENESS & EXPRESSIVENESS, to model the system from architectural view point as well as from the unit view point. (k) CONSISTENCY, This property requires that there should be no contradiction between models [4]. (l) MODEL REUSABILITY, the ability of any component to be re-used by other system with minor or even no modifications. (m) ABSTRACTION & MODULARITY, abstraction deals with the ability of the AOM to produce models at various levels of details

Modularity is the property to divide the system in small manageable chunks.

Technique Related Criteria: This criteria deal with assessing the methodology's techniques to perform development steps and/or to produce models and notational components.

- (a) AVAILABILITY OF TECHNIQUES & HEURISTICS, This is the property of an AOM to provide techniques to perform each process step. Techniques to produce each model and notational components.
- (b) TECHNIQUE USABILITY, AOM should provide a systematic structure to be followed in order to develop a system model.
- (c) EASE OF UNDERSTANDING, The notations provided by the AOM must be easy to learn & remember by different type of users [14]. This requires inclusion of the symbols and notations which are familiar to the users.

Process Related Criteria: This criteria looks at the applicability of the AOM, the steps provided for development process and the development approach followed by the AOM.

(a) DEVELOPMENT LIFE CYCLE, This criteria state about the development context supported by the AOM. Whether it supports waterfall model, prototype model, iterative enhancement model etc.

(b) DEVELOPMENT PROCESS STEPS, This criterion evaluates the tasks and activities specified by the AOM for the development process.

(c) VERIFICATION & VALIDATION SUPPORT, Are we building the system right? Are we building the right system? Both the questions must be answered in order to have a clear idea about correctness of the developed models and specified requirements.

(d) REFINABILITY, a simplified sequence of steps must be provided by the methodology to add new details in the existing model. Refinement allows the developers to make necessary changes at gradual stages of design development in an easy and simplified way [15].

Support Features Related Criteria: These are "add on" features provided by any methodology. This criterion assesses various supplementary features provided by any AOM. It includes CASE tools to support dynamic and open systems which allow dynamic addition and removal of the agents. Support for mobile agents and conjunction of conventional objects in the MAS are also included in the supportive features.

SOFTWARE & METHODOLOGICAL SUPPORT, This criterion assesses availability of various development support tools like CASE tools and libraries to develop MAS. (b) OPEN SYSTEM DEVELOPMENT SUPPORT, Multi-agent systems are dynamic in nature. Various agents interact with each other to perform their goal. In a dynamic open system, agents can be added or removed in and from the system at any point of time. This criterion assesses support provided by an AOM to develop open agent based system.

EVALUATION RESULTS

We have compared features of five AOMs GAIA, MaSE, PROMETHEUS, TROPOS & MESSAGE using the above mentioned feature analysis framework. We have done a primary survey also using a questionnaire among the users of these methodologies along with presenting the claims made by the developer of particular AOMs and we have also included our experience regarding the same. The primary research work is done to sideline the chances of biasing towards a particular AOM. The questionnaire consists of twenty one questions in total divided into four sections and is based upon the modified comparison framework discussed in earlier section. The Combined result of the questionnaire and our observation is presented in the subsequent sections. The results are mentioned on an abstract scale of H, M, L, N and X where H stands for High, M for Medium, L for Low, N for Not Available, X for can't say. The results are as under.

AUTONOMY: From the comparison provided in table 1, we can analyze that almost all the five AOMs are having high ratings with respect to this property. This due to the fact that all of these AOMs have constructs available to implement the autonomous property. For instance TROPOS has plan diagrams, PROMETHEUS has plan descriptor and MaSE consist of task state diagram. These constructs are good enough to implement agent plans and reasoning rules in order to implement the autonomous property for an agent based system. **REACTIVITY:** As shown in the comparison table 1, again the rating is on high to medium scale. TROPOS has the Actor Diagrams and PROMETHEUS has the Agent Class Descriptor to implement the reactive behavior of the agent based system. **CONCURRENCY:** PROMETHEUS and GAIA are considered to be having very low facility available for concurrency; the strongest AOM for concurrency implementation is MaSE due to availability of constructs like Task State Diagrams & Communication Class Diagram. These diagrams are helpful to define coordination protocols between two agents and thus achieving the important attribute of Concurrency. **PRO-ACTIVENESS:** Except MESSAGE, we got approximately similar response for all five AOMs regarding this criterion. PROMETHEUS seems to be the best amongst all due to availability of Action Descriptors. Using this construct the agents can be modeled in the way such that they respond to the goals evolving over time in the environment. **ENVIRONMENT BELIEF:** From the survey and our own observations, we found that GAIA and PROMETHEUS provide clear constructs for Environment Belief.

They allow modeling of agents in the way such that agents can capture information from the environment and appropriate processing can be done.

GAIA provides Environmental Model and PROMETHEUS provides System Overview Diagram which is also known as the Environmental Model. **COOPERATIVE BEHAVIOR:** Agents cannot exist in a vacuum. Agents need support of other agents and they have to provide support also. Agents can delegate their task to other agents, can negotiate with other agents and can work in a shared way also. Acquaintance of one agent with other can be easily modeled in GAIA as they provide acquaintance model for the same,

TROPOS has Sequence/Collaboration diagram, PROMETHEUS provides Interaction Diagram, MaSE provides Agent Class diagram and MESSAGE provides Organization Model. From our own experience, we found that the constructs provided by TROPOS & MESSAGE are capable of modeling any kind of agent acquaintance.

COMMUNICATION ABILITY: Agents can directly, indirectly, synchronously and asynchronously interact with each other. From the survey and our own experience we felt that almost all five AOMs provide satisfactory constructs for the variety of communication modes. **ACP (AGENT COMMUNICATION PROTOCOL):** All five AOMs in consideration provide good constructs with some minor limitations. GAIA provides Interaction model for the same but with the limitation that contents of exchanged message between agents cannot be defined in GAIA model. TROPOS provides Sequence Diagrams, PROMETHEUS provides Interaction Protocol Diagrams, MaSE provides Communication Class Diagram and MESSAGE provides Interaction Model.

ACL (AGENT COMMUNICATION LANGUAGE): An ACL provides means to agents to exchange information and knowledge with other agents. All five AOM's basic communication language is based upon the "*speech act*" where not only contents but intentions and actions also matters. Both KQML and FIPA-ACL are supported by all the AOMs under consideration. In table 1, from the survey and from our observation, it can be seen that the ratings are medium to high for all the AOMs. **COMPLETENESS & EXPRESSIVENESS:** From the survey and from our own experience we analyzed that all five AOM provide fairly good & well defined symbols and notations. GAIA, MaSE, TROPOS, MESSAGE and PROMETHEUS provide adequate constructs to completely express complex and dynamic system. One of the team members does not find syntax and symbols of GAIA and TROPOS satisfactory to completely model MAS, but from our own experience we felt that the TROPOS may be the possible candidate for the claim made by the team, as it only provides some help during detailed design. For GAIA, we felt that symbols and notations are quite satisfactory. **CONSISTENCY:** This attribute differs a lot from one AOM to another. Consistency requires there must be a consistent relationship between modeling and design i.e. inter-model and intra-model consistency. The Prometheus Design tool (PDT) in PROMETHEUS and agent-Tool in MaSE provide enough support for the design and model consistency check. MESSAGE only provide limited consistency check in the form drawing diagrams. GAIA and TROPOS do not provide support for consistency check at all.

MODEL REUSABILITY: Either the teams are not sure about availability of this criterion or they felt it at very low level. This is due to the fact that none of the AOM under consideration provides any explicit construct to implement model reusability. Though TROPOS, PROMETHEUS and MaSE claims for model reusability but no formal guidelines are available to design reusable components in any of the AOM.

ABSTRACTION & MODULARITY: Almost all the five AOMs are having medium to high ratings in this criteria. The Agent model in GAIA, Agent/Role model in MESSAGE and Agent Class Diagram in MaSE etc. provides sufficient constructs for achieving abstraction and modularity while modeling or designing any MAS.

Technique Related Criteria constitutes:

AVAILABILITY OF TECHNIQUES & HEURISTICS: It deals with the availability of clearly defined techniques to perform each process step and to produce each model and notational components. It is observed that all five AOMs under consideration provide fairly good support at each step either implicitly or explicitly. Right from identifying system tasks to system deployment, all AOMs provide various constructs for the same. For instance GAIA provides Role Model for Identifying system tasks in an implicit way and Agent Model for specifying agent classes.

Other techniques provided by GAIA are Interaction Model, Service Model, and Environmental Model etc. being used at different steps. TROPOS has Actor Diagram, Plan Diagram and Sequence Diagram. PROMETHEUS has Goal Diagram, Agent Class Descriptor, Interaction Diagrams & Protocols and Capability Diagram etc. MaSE provides Goal Hierarchy Diagram, Agent Class Diagram, Communication Class Diagram and Deployment Diagram etc. MESSAGE has Task Model, System Architecture Diagram and Organizational Model etc.

TECHNIQUE USABILITY: As discussed earlier, MaSE, PROMETHEUS & MESSAGE provides have integrated tool support to draw diagrams and check model & design consistency. TROPOS is an exception where no such kind of facility is available, though the developers claim that the notations and symbols are fairly easy to understand. Still, we have maintained low ratings for the TROPOS regarding usability criteria.

EASE OF UNDERSTANDING: “*Ease of learning*” any AOM is concerned with many criteria like unambiguous syntax and semantics, clear and expressiveness nature etc. All the AOMs under consideration are having medium to high ratings for this criteria.

DEVELOPMENT LIFE CYCLE: It deals with the nature of development life cycle. GAIA supports iterative development within each phase but sequential between phases, TROPOS is iterative & incremental, PROMETHEUS & MaSE are iterative across all phases and MESSAGE follows RUP life cycle. So rather than giving the ratings to a particular AOM, we have mentioned the nature of development life cycle being followed by the AOM.

DEVELOPMENT PROCESS STEPS: It deals with the clear separation of phases/steps of development process. We observed that except MaSE, no AOM provides clear and explicit steps for Testing, Debugging, Deployment and Maintenance. Even Implementation step is also not provided at satisfactory level by all of the AOMs. All the AOMs though provide good support for Requirement Analysis, Architectural Design and Detailed Design. Considerable work still needs to be done in the later phases the development process of agent oriented system.

VERIFICATION & VALIDATION SUPPORT: V & V is an essential activity to achieve quality of an agent based system. Except MaSE and PROMETHEUS, no AOM under consideration provides support for verification and validation. MESSAGE has kept this feature as future development.

REFINABILITY: Again, all the AOMs got medium to high ratings for this feature. Developers are free to roam in different process steps to enhance the details in the existing model.

CONCLUSION & CRITICAL DISCUSSION

There is very strong demand of developing complex software systems for industrial and general applications. Agents seem to be the best solution for the same. Attributes like autonomy, reactivity, pro-activeness etc provide a fantastic platform to design and develop complex systems. The need is to compare and evaluate various advantages and disadvantages of various AOMs. As various AOMs has been proposed and discussed in the literature, the aim of this paper is to provide an unbiased

comparison of different methodologies using a modified feature analysis framework along with the primary survey technique. We have carried out comparison of five selected AOMs. The purpose of the comparative study is not to prove one AOM superior or inferior over another but to figure out strengths, weakness, domain applicability, similarities and dissimilarities as compared to each other.

The aim of modifying the feature analysis framework is to compare the software engineering attributes provided by a particular AOM in terms of classical software engineering paradigms, object oriented paradigms and those which are specific to agent based development.

Another reason is to provide a multistage comparative analysis to cover all significant software engineering criteria. This is required to enhance the overall software development experience of the developing team to develop a complex system using agent oriented software engineering paradigms. We have incorporated four basic paradigms which an AOM supports. These are model related criteria, technique related criteria, process related criteria and supportive features related criteria. In addition to the qualitative analysis of the AOMs we have also used the primary research technique of questionnaire, where we have collected views of the users of agent oriented methodologies. This is required to eliminate the chances of any kind of biasing in the comparative analysis.

From the results obtained we can see that all five AOMs provide reasonably good support for the features like pro-activeness, autonomy, reactivity etc. required for developing an agent based application. All five AOMs are also considered as pure agent oriented methodologies rather than merely the extension of object oriented methodologies. All five AOMs have clear and understandable notations to model and develop the agent based system. Along with some good similarities the AOMs under consideration have some dissimilarity also. TROPOS seems to be difficult to use and understand. MaSE

GAIA seems to be providing less support in terms of expressiveness. Only PROMETHEUS & MaSE provides tool support to check consistency between models. MESSAGE and GAIA does not provide support for detailed design. On one hand PROMETHEUS & MaSE provide good heuristics support for architectural and detailed design, on the other hand MESSAGE provide no heuristic support for the same. In addition to the individual pros and cons, all AOMs share some good and bad points.

None of the five AOMs under consideration provide explicit feature to design team work in multi agent system. Environmental modeling constructs are also not fully provided in the all five AOMs. Implementation, Testing, Debugging and maintenance phases are either poorly defined or not defined at all in all five AOMs.

Various other factors which are important in industrial terms like project management techniques, software quality assurance techniques, cost and effort estimation etc. are also not included in any of the AOM. An AOM which can be considered as a benchmark for developing an agent oriented system with all relevant feature support is still to be developed; the need is to use the positive features of

each AOM and develops a perfect AOM rather then moving in separate and scattered directions.

- [15] Nicholas R. Jennings."An agent-based approach for building complex software systems." Communications of the ACM, 44(4):35–41, 2001.

REFERENCES

- [1] David Law, "Methods for Comparing Methods: Techniques in Software Development". NCC Publications, 1988.
- [2] Barbara Kitchenham, "DESMET:a method for evaluating software engineering methods and tools." Technical Report TR96-09, University of Keele, U.K., August 1996.
- [3] Tran, Q.N., Low, G., Williams, M.A., "A Feature Analysis Framework for Evaluating Multiagent System Development Methodologies." In Zhong, N., Ras, Z.W., Tsumoto, S., Suzuki, E. (eds): Foundations of Intelligent Systems – Proc. of the 14th Int. Symposium on Methodologies for Intelligent Systems ISMIS'03 (2003) pp 613-617.
- [4] Wood B., Pethia R., Gold L.R., and Firth R."A guide to the assessment of software development methods. Technical Report 88-TR-8", Software Engineering Institute, Carnegie-Mellon University, Pittsburgh, PA, 1988.
- [5] Jayaratna N., "Understanding and evaluating methodologies – NIMSAD a systematic framework". Maidenhead, UK: McGraw-Hill. 1994
- [6] Olle T.W., Sol H.G., & Tully, C.J., "Information systems design methodologies – A feature analysis." Amsterdam: Elsevier Science Publishers. 1983
- [7] O'Malley S.A. & DeLoach S.A."Determining when to use an agent oriented software engineering paradigm". In Proceedings of the 2nd International Workshop on Agent-Oriented Software Engineering (AOSE 2001), Montreal, Canada, (pp. 188-205). Springer-Verlag. May 29, 2001.
- [8] Cernuzzi L. & Rossi G., "Evaluation of agent-oriented modeling methods". In Proceedings of the OOPSLA Workshop on Agent-Oriented Methodologies, Seattle, (pp. 21-33). University of Technology, Sydney: Centre for Object Technology Applications and Research. November 4-8 , 2002
- [9] Sabas A., Badri M., & Delisle, S., "A multidimensional framework for the evaluation of multi-agent system methodologies". In Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, (pp. 211-216). Orlando: International Institute of Informatics and Systemics. July 14-18, 2002
- [10] Frank U."Evaluating modeling languages:relevant issues, epistemological challenges and a preliminary research framework. Technical Report 15", Arbeitsberichte des Instituts fuer Wirtschaftsinformatik (Universitt Koblenz-Landau), 1998.
- [11] Avison D. and Fitzgerald G., "Information Systems Development: Methodologies, Techniques and Tools." McGraw-Hill, New York, 2nd edition, 1995, pp 452.
- [12] Jacques Ferber,"Multi-agent Systems: An Introduction to Distributed Artificial Intelligence." Addison-Wesley, 1999 pp 78-82.
- [13] Munindar P. Singh, "Agent communication languages: Rethinking the principles." IEEE Computer, 31(12):40–47, December 1998.
- [14] Rumbaugh J. "Notation notes: Principles for choosing notation." Journal of Object-Oriented Programming 15 (JOOP), 8(10):11–14, May 1996.

Suitability of Addition-Composition Fully Homomorphic Encryption Scheme for Securing Data in Cloud Computing

Richard Omollo¹ and George Raburu²

^{1,2}Department of Computer Science and Software Engineering

Jaramogi Oginga Odinga University of Science and Technology

P.O. Box 210-40601, Bondo-Kenya.

Email: comolor@hotmail.com¹ and graburu@hotmail.com²

Abstract: Cloud computing is a technological paradigm that enables the consumer to enjoy the benefits of computing services and applications without necessarily worrying about the investment and maintenance costs. This paper focuses on the applicability of a new fully homomorphic encryption scheme (FHE) in solving data security in cloud computing. Different types of existing homomorphic encryption schemes, including both partial and fully homomorphic encryption schemes are reviewed. The study was aimed at constructing a fully homomorphic encryption scheme that lessens the computational strain on the computing assets as compared to Gentry's contribution on partial homomorphic encryption schemes where he constructed homomorphic encryption based on ideal lattices using both additive and multiplicative Homomorphisms. In this study both addition and composition operations implementing a fully homomorphic encryption scheme that secures data within cloud computing is used. The work is founded on mathematical theory that is translated into an algorithm implementable in JAVA. The work was tested by a single computing hardware to ascertain its suitability. The newly developed FHE scheme posted better results that confirmed its suitability for data security in cloud computing.

Keywords: Cloud Computing, Data Security, Fully Homomorphic Encryption Schemes, Addition-Composition

I. INTRODUCTION

Cloud computing paradigm traces back to 1960s suggestion by Professor John McCarthy who proposed the concept of *utility computing*. Utility computing is about computer time-sharing technology leading to a future where computing power and even specific applications provided through utility-type business model [1]. This has been the motivation behind cloud computing, a compacted term within technology circles, which adopts a distributed computing style of integrating web services and data centers [2]. However, this paradigm can be easily extended to grid computing, which is a form of distributed computing that implements a virtual supercomputer composed of a cluster of network or internetworked computing dedicated towards a demanding task.

Cloud computing therefore is an internet based system that provides the technology of outsourcing computational needs away from the data owners. It provides a way of storing and accessing cloud data from anywhere by just connecting to the cloud application using the internet [3]. The cloud users have the choice of settling on cloud services of their interest so as to store and transact their data in a remote data server [4]. This data can be accessed and managed through cloud services provided by the cloud providers chosen by the user.

Despite the flexibility and convenience that comes with cloud computing, the security of cloud data has been one of the greatest challenge to its deployment [5]. In order to address this, cryptography is one approach that guarantees safety of cloud resources [6] since it has the capability of hiding data from unauthorized access. There has been a proposed solution to address issues of data security on cloud computing with the adoption of homomorphic encryption scheme supporting only limited mathematical operations.

Fig. 1 below is illustrates an implementation of an ideal cloud service that supports homomorphic encryption in the clouds.

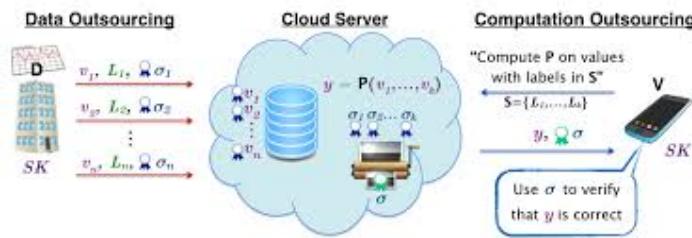


Fig. 1 Homomorphic Encryption in the Clouds

In order to strengthen homomorphic encryptions in clouds a Fully Homomorphic Encryption Scheme (FHE) has been developed that involves more than one operation as proposed by Craig Gentry [7]. More contributions have been made on Gentry's work but still far off from ensuring effective practical applications. As much as FHE schemes have proved versatile, the computing resource demand due to increasing ciphertext size and slowed encryption speed still compromises the implementation. Based on this critical need a new fully homomorphic scheme that is still versatile but also faster in encryption or decryption and guarantees ciphertext size of higher value was developed. This new scheme addresses the existing gap where previously developed schemes strain computing resources such as storage and bandwidth consumption. The use of both additive and composition mathematical operations is adopted in this study to address data security in cloud computing.

II. LITERATURE REVIEW AND RELATED STUDY

2.1 CLOUD COMPUTING

A more functional understanding of Cloud Computing is tied to the concept of the term CLOUD. It [8] derived five logical tautology and concludes that a cloud structure is a six-tuple (space, time, directed graph, set of states, transition function, and initial state) that satisfies the five axioms namely; Common, Location-independent, Online, Utility, and on-Demand (C.L.O.U.D). Cloud computing can therefore be understood to have five distinct essential characteristics: On-Demand Self-Service, Broad Network Access, Resource Pooling, Rapid Elasticity, and Measured Service [9]. Likewise from a hardware point of view, there are three new aspects in the paradigm of cloud computing [10].

2.2 ARCHITECTURE OF CLOUD COMPUTING

The design of Cloud computing architecture encompasses the front-end (a visible interface that cloud users encounter through web-enabled client devices) and the back-end (comprising resources required to deliver cloud computing services). The back-end consists of the bare metal servers, data storage compartments, virtual machines, implemented security mechanism, and cloud service(s) with conformity with a particular deployment model. It is the primary responsibility of the back-end to integrate built-in security mechanisms, traffic management control, and protocols. In the bare metal has the operating system referred to as hypervisor which makes use of well-defined protocols enabling efficient concurrent access onto the virtual machines.

The cloud computing architecture functions in such a way that all applications are controlled and served by a central cloud server with replicated data remotely in the cloud configuration to create limitless efficiency and possibilities.

Fig. 2 below shows a diagrammatic summary of cloud computing architecture.

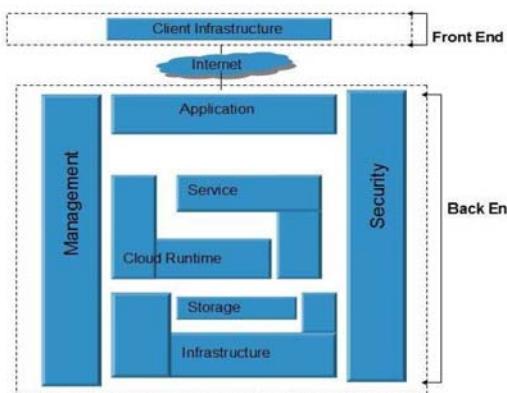


Fig. 2 Cloud Computing Structure

Cloud services are based on five principal characteristics that demonstrate their relation to, and differences from, traditional computing approaches [11]. These characteristics are summarized in Table 1:

Table 1 Cloud Principal Characteristics

Characteristics	Descriptions
Abstraction of Infrastructure	The computation, network and storage infrastructure resources are abstracted from the application and information resources as a function of service delivery. This abstraction is generally provided by means of high levels of virtualization at the chipset and operating system levels or enabled at the higher levels by heavily customized file systems, operating systems or communication protocols.
Resource Democratization	The infrastructure, applications, or information provides the capability for pooled resources to be made available and accessible to anyone or anything authorized to utilize them using standardized methods
Service Oriented Architecture	Utilization of cloud components in whole or part, alone or with integration, provides a services oriented architecture where resources may be accessed and utilized in a standard way.
Elasticity or Dynamism	The on-demand model of cloud provisioning coupled with high levels of automation, virtualization, and ubiquitous, reliable and high-speed connectivity provides for the capability to rapidly expand or contract resource allocation to service definition and requirements using a self-service model that scales to as-needed capacity.
Utility Model of Consumption and Allocation.	The abstracted, democratized, service-oriented and elastic nature of cloud combined with tight automation, orchestration, provisioning and self-service allows for dynamic allocation of resources based on any number of governing input parameters. Given the visibility at an atomic level, the consumption of resources can then be used to provide a metered utility cost and usage model.

2.3 CLOUD COMPUTING SERVICE DELIVERY MODELS

In the delivery models in cloud computing, there are three service models that commonly established and formalized: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The three archetypal models and the derivative combinations thereof generally describe cloud computing service delivery. The three individual models are often referred to as the “SPI MODEL” [11]. There are other specialized variations of the three main service delivery models based on distinct combination of IT resources.

2.3.1 Software as a Service (SaaS): This is a service delivery model where applications are available to the cloud users over the network. The cloud providers take the responsibility of hosting these applications used by the consumers under software distribution model. The model is characterized by application delivery from 1:M model, that is, single-instance, multitenant architecture as opposed to traditional 1:1 model, and also associated with pay-as-you-go subscription licensing model. The distinguishing bit of SaaS from other service delivery model is that it was designed to work with web browsers. The main advantages with this model include; automated updates and patch management services, data compatibility across the enterprise and global accessibility of software of interest by the cloud users. The design is made in such a way that the cloud consumer depends on the service provider in the implementation of security e.g. the provider has to ensure that each cloud user do not access each other’s private data [12]. The cloud consumer substitute new software applications with the old one focusing on enhancing the security functionalities provided by legacy application and achieving a successful data migration [13].

2.3.2 Platform as a Service (PaaS): This is service delivery model where all facilities required the support of the complete life cycle of building and delivery of web applications and services entirely available on the Internet. It is an outgrowth of SaaS model but uses web-based development unlike with SaaS where specific operating systems instance is favoured. It encapsulate a layer of software and provide it as a service used to build higher levels services with two perspective of the producer or consumer of services: producing a platform by integrating an operating system, and an encapsulated service presented to them through Application Programming Interface (API) [14]. The advantage with this model is that it provides the entire infrastructure needed by allowing users to focus on innovation and not the complex infrastructure. A secure PaaS cloud can only be achieved through thoroughly inspecting the conceptual security solutions, architectural and software design, implementation and service provisioning [15]. Also for PaaS model, applications are deployed without the necessity of purchasing and maintaining the hardware and software thereby depending on a secure browser. PaaS application security includes the security of application deployed on PaaS as well as the PaaS platform security itself and it is therefore the responsibility of the PaaS provider to protect the runtime engine which runs the client applications [16]. This deployment model has its products available with different development stacks.

2.3.3 Infrastructure as a Service (IaaS): This is service delivery model where cloud service providers manage the transition and hosting of selected applications on their infrastructure. In this model, the computing services are provided on a virtualized environment with cloud users maintaining ownership and management of the applications while hosting operations and infrastructure management are off-loaded to the cloud provider. The benefits that comes with this service delivery model includes availability of resources on demand from any location 24x7 when required by the cloud users. The physical security of cloud users' data and any chance of system failure is smoothly handled by the cloud provider. In IaaS, the security varies on deployment model adopted in its implementation. For instance, in private cloud there is control over solutions from top to bottom while in public cloud, there is control on the VMS and services running on the VMS.

The three main service delivery models can be captured in the Fig. 3.



Fig. 3 Security Checklist for SAAS, PAAS, and IAAS

2.4 CLOUD COMPUTING DEPLOYMENT AND CONSUMPTION MODELS

Regardless of the cloud service delivery model utilized, there are five models that can be used to deploy cloud services [11]. For specific organization, cloud integrators play a vital role in determining the right cloud path to be adopted. The models are as listed below:

2.4.1 Private Cloud: Is provisioned exclusive use by a single organization with multi-tenants resource access. It may be owned, managed and operated by the organization, a third party or some combination of them.

2.4.2 Public Cloud: Is provisioned for open use of resources dynamically by the general public. It may be owned, managed and operated by one or more organizations who sells the cloud services.

2.4.3 Community Cloud: Is provisioned for exclusive use by a specific community of clients from organizations that share common interests like mission, security requirements, policy or compliance considerations. It is owned, managed and operated by one or more organization in a community.

2.4.4 Hybrid Cloud: Is composed of two or more distinct cloud infrastructures characteristics (private, public or community) that remain unique entities, but are bounded together by standardized or proprietary technology that enables data and application portability. It is owned, managed and operated by both cloud infrastructures incorporated.

2.4.5 Virtual Private Cloud: Is allocated within a public cloud environment provisioned for a certain level of isolation between the different organizations sharing the cloud resources. It is viewed as a hybrid model of cloud infrastructure in which a private cloud solution is provided within a public cloud infrastructure. Its ownership, management and operations is assumed by the public cloud infrastructure. The Table 2 below summarizes cloud deployment models, with each viewed on access security and infrastructure where it is hosted.

Table 2 Summary of Cloud Deployment Models

Model	Infrastructure		Access	
	Location	Ownership	Trusted	Untrusted
Private	On Premise	Organization	Yes	No
Public	Off Premise	Third Party	No	Yes
Community	Off premise	Third Party	Yes	No
Hybrid	On and Off Premise	Organization and Third Party	Yes	Yes
Virtual Private	Off Premise	Third Party	Yes	Yes

The security issues of the five cloud computing deployment are illustrated in Table 3.

Table 3 Cloud Deployment Models and their Security Issues

Model	Security Issues	Control Issues
Private	Most secure	Most control
Public	Least secure, multi-tenancy, and transfer over the Internet	Least control
Community	Least Secure	Least Control
Hybrid	Control of security between private and public clouds	Least control
Virtual Private	Most secure	Most Control

2.5 CLOUD COMPUTING SECURITY

In cloud architecture the security components and services must be transparent and generic: transparent since there is need for automatic application without much human intervention, and generic to ensure adjustability on the part of clients, requirements, applications and required services. A functional cloud computing security architecture is composed of Security Access Points that provides front-end security services, Security infrastructure servers that manage all cloud stored data registered in the cloud, and secure client for cloud standard clients stations extended with some security components.

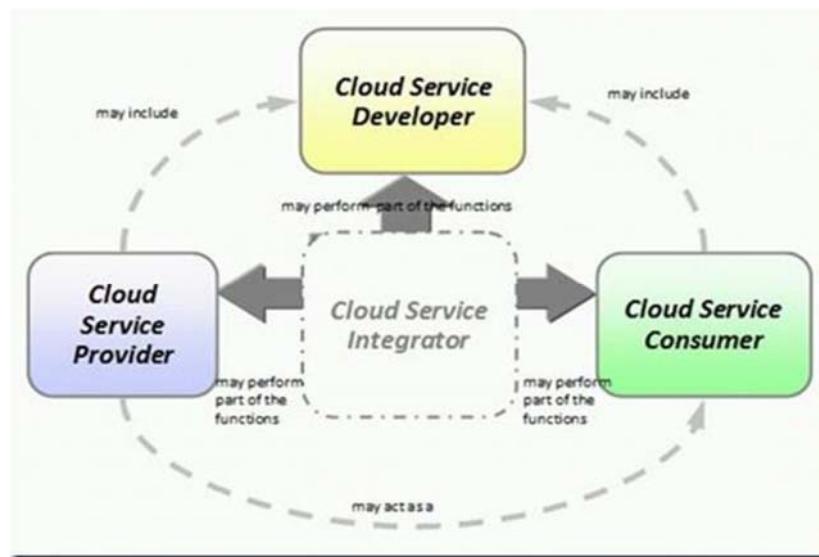
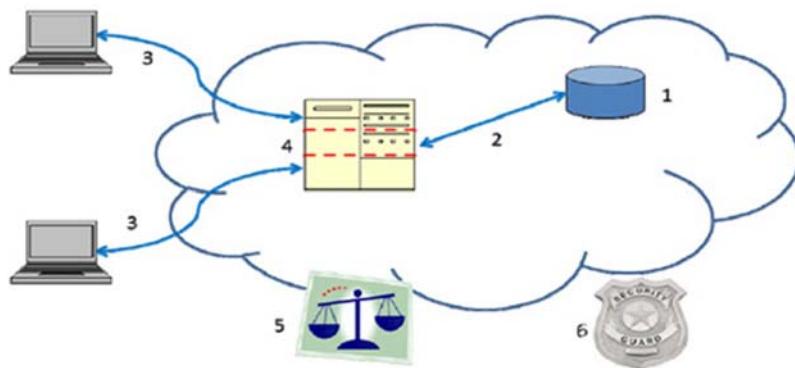


Fig. 4 Cloud Computing Security Infrastructure

The multi-tenancy model and the pooled computing resources has introduced new security challenges such as shared resources on the same physical machines inviting unexpected side channels between machine resources and a regular resource [17]. Security being an important component of cloud computing deployment, it is necessary to address it to enable cloud consumers enjoy its benefits uninterrupted [14].

Cloud computing architecture has numerous security issues as it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Therefore this means that security issues affecting many of these systems and technologies are applicable to cloud computing. Take for instance the network that interconnects the systems in a cloud has to be secure. Furthermore, virtualization paradigm in cloud computing leads to several security concerns e.g. mapping the virtual machines to the physical machines has to be carried out securely. Data security involves encrypting the data as well as ensuring that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms have to be secure. Finally, data mining techniques may be applicable for malware detection in the clouds – an approach which is usually adopted in intrusion detection systems (IDSs) ([18], [19], [20], [21], and [22]). Fig. 5 illustrates the six specific areas of the cloud computing environment where equipment and software require substantial security attention [23].



(1) security of data at rest, (2) security of data in transit, (3) authentication of users or applications or processes, (4) robust separation between data belonging to different customers, (5) cloud legal and regulatory issues, and (6) incident response.

Fig. 5 Areas of Security Concerns in Cloud Computing

For securing data at rest, cryptographic encryption mechanisms are certainly the best options. The hard drive manufacturers are now shipping self-encrypting drives that implement trusted storage standards of the trusted computing group [23]. These self-encrypting drives build encryption hardware into the drive, providing automated encryption with minimal cost or performance impact. Although software encryption can also be used for protecting data, it makes the process slower and less secure since it may be possible for an adversary to steal the encryption key from the machine without being detected. Encryption is the best option for securing data in transit as well. In addition, authentication and integrity protection mechanisms ensure that data only goes where the customer wants it to go and it is not modified in transit. Strong authentication is a mandatory requirement for any cloud deployment. User authentication is the primary basis for access control. In the cloud environment, authentication and access control are more important than ever since the cloud and all of its data are accessible to anyone over the Internet. The trusted computing group's (TCG's) IF-MAP standard allows for real-time communication between a cloud service provider and the customer about authorized users and other security issues. When a user's access privilege is revoked or reassigned, the customer's identity management system can notify the cloud provider in real-time so that the user's cloud access can be modified or revoked within a very short span of time [24]. The problem is that the cloud service being a web application, it contains data remembrance or persistence remains an issue due to the replication and distribution of data even after user left a cloud provider. This can be corrected by use homomorphic encryption schemes to protect data on cloud [25].

2.6 CHALLENGES AND SOLUTIONS TO DATA SECURITY IN CLOUD COMPUTING

The security of Cloud Computing is the major concern to be addressed since if the security measures are not provided properly for data operations and transmissions then data is at high risk [26]. This is accelerated by the fact that cloud computing do provide facilities for a group of cloud users to access the stored data thus there is a possibility of having high data risk. Therefore it is necessary to say that strongest security measures need to be implemented by identifying security challenges and solutions to handle these challenges [27].

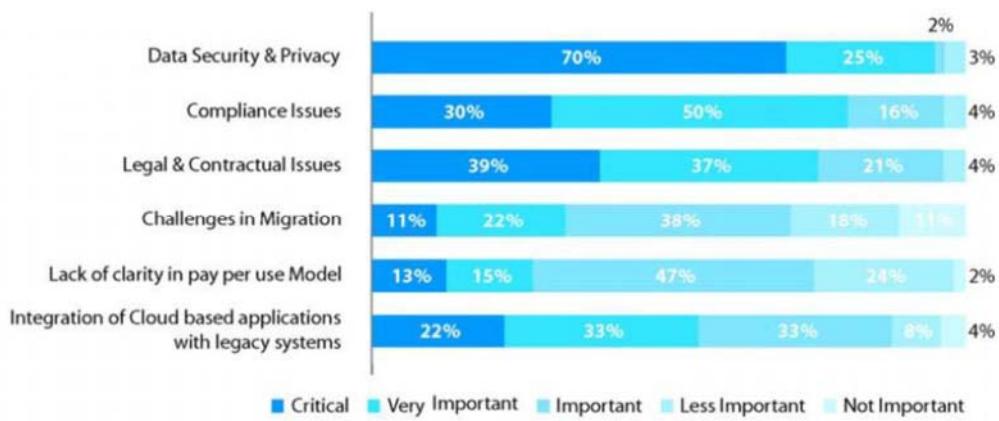


Fig. 6 Factors Influencing Cloud Computing Adoption [27]

From Fig. 6 illustrates the importance of Data Security and Privacy are most important in the adoption of cloud computing.

2.7 HOMOMORPHIC ENCRYPTION SCHEMES

The term *homomorphism* is borrowed from algebra meaning a structure-preserving map between two algebraic structures of the same type. The concept of homomorphism has been employed in explaining *homomorphic*, which means in literal sense same effect for different expression. Homomorphic encryption enables computation to be performed on the encrypted data without requiring the secret key. This makes it a better approach to enhance security of sensitive data stored and manipulated on untrusted storage systems [28]. Therefore an efficient and fully homomorphic cryptosystem have great practical benefits to outsourcing private computations.

The concept of homomorphic encryption was introduced in 1978 by Rivest, Adleman and Dertouzos after the development of RSA algorithm [29], where they highlighted four possible encryption functions which include RSA as additive and multiplicative privacy homomorphism. Since then, many encryption schemes have been developed but they were either using addition or multiplication homomorphic computations. Later, an encryption scheme was developed [30] that perform unlimited number of addition operations but single multiplication.

The realization of fully homomorphic encryption scheme appeared elusive for more than thirty years until in 2009 when Craig Gentry from IBM implemented the first version that could perform many additions and multiplications using ideal lattices and bootstrapping technique [7]. He based his construction on ideal lattices starting with constructing somewhat homomorphic encryption scheme, limited to evaluating low-degree polynomials over encrypted data. He then modified it to make it bootstrappable hence capable of evaluating its own decryption circuit with at least one more operation. Even though Gentry's scheme was able to demonstrate the possibility of fully homomorphic encryption scheme, there was still need to improve on its complexity, efficiency and performance. For instance, he estimated that building a circuit to execute an encrypted Google search with encrypted keywords would multiply the current computing time by one trillion. More research have been done with the motivation to improve on Gentry's work. These are referred to as second generation of homomorphic encryption schemes and have their security based on the hardness of the Learning with errors problem apart from the LTV scheme [31] whose security is based on a variant of the NTRU computational problem.

2.8 NOISE GROWTH, ATTACKS AND EFFICIENCY IN FULLY HOMOMORPHIC ENCRYPTION SCHEMES

In the context of homomorphic encryption schemes, *noise* is often referred to as a small term added into the ciphertext while encrypting. It depends with the application. It may be a small integer if the scheme is based on integers or a small polynomial if the scheme is based on polynomial. This noise plays a significant role in the encryption process in that it is responsible for the security of the cryptosystem. However, when it comes to decipherment in the same case, the decryption function gets affected by the noise term especially if it is greater than a certain maximum value. Homomorphic operation usually increases the noise and it is advisable to limit the number of operations in order to control it. For instance in the case of fully homomorphic encryption schemes over integers [32], to encrypt a bit m into an integer c with p being the private key and sample values of q and r are considered using the equation:

$$c = pq + 2r + m \dots \text{Eqn. 1}$$

For decrypting the value c , $c \bmod p$ gives $2r+m$ only if $2r+m$ is smaller than p thus $\bmod 2$ is reduced to end up with m . Also it is important to note that the decryption won't work if $2r$ is bigger than p , thus makes $2r$ to be the noise term in this case. This is explained further when the homomorphic property of the scheme is considered. For example in a case where there are two ciphertexts c_1 and c_2 , then applying the addition operation will end up with $c_1 + c_2 = p(q_1 + q_2) + 2(r_1 + r_2) + (m_1 + m_2)$ hence increasing the noise. The multiplication operation further complicates it and therefore with each operation increasing the noise, it is advisable to set an acceptable maximum value for the noise by limiting the number of operations.

The significance of noise term in the construction of fully homomorphic encryption scheme is an important factor to be considered. Practical constructions should consider the reduction of noise term as it complicates its decryption function. For instance [33] presented a fully homomorphic encryption scheme that he claimed to be efficient for practical applications. He avoided bootstrapping and modulus switching as noise reduction approaches that were evident in previous FHE schemes but included it in the algorithm. According to [34], they proposed two fully homomorphic encryption schemes (symmetric and asymmetric) which were noiseless in the sense there was no *noise* factor contained in the ciphertexts.

Since the introduction of privacy homomorphism by [29], all schemes have been insecure until Gentry introduced the first fully homomorphic encryption scheme based on ideal lattices [35]. More improvements have been built on Gentry [7] with the simplest done by [32] that considered known attacks on the Approximate Greatest Common Divisor problem for two numbers (x_0, x_1) and many numbers (x_0, \dots, x_t) . These attacks were mainly explored with a view of solving the approximate greatest common divisor problem, which is the secret key, p .

In the study of the attack on fully homomorphic encryption scheme, lattice attack based on the public key is considered. For instance in the security of FHE scheme, it was proved by [32] that it is equivalent to solve the approximate greatest common divisor problem. Consider a list of approximate multiples of p :

$$\{x_i = qp + r_i : q_i \in \mathbb{Z} \cap [0, 2^t/p), r_i \in \mathbb{Z} \cap (-2^t, 2^t)\}_{i=0}^t \quad \dots \dots \dots \text{Eqn. 2}$$

According to FHE scheme, it is already known that an arbitrary ciphertext c has a general form: $c = qp + 2r + m$. In order to execute the attack here, it is very simple, that is, how to remove qp in a ciphertext c by adding small noise value. When completing this, it is easy to recover the plaintext m bit in c . The success of this calls for applying Diophantine Inequality Equation (DIE) problem [36].

The study has explored the noise term and its significant contribution to the performance of the FHE scheme especially where the noise factor impact heavily on the decryption function of the scheme. Thereafter the study has also explored the principle behind the attack in FHE scheme where the application of DIE problem has been proposed. Both the noise term and attacks impacts proportionally on the efficiency of the FHE schemes.

After the breakthrough in [7], the first attempted implementation of fully homomorphic encryption was the [37] implementation based Gentry's original cryptosystem where they reported timing of about thirty minutes per basic bit operation. In further evaluation of the same, an implementation from a variant of the [32] cryptosystem, reported evaluation of a complex circuit in thirty-six hours. While using the packed-ciphertext techniques, that implementation could evaluate the same circuit on fifty-four different inputs in the same thirty-six hours, yielding amortized time of roughly forty minutes per input. The AES-encryption circuit was then adopted as a benchmark in several subsequent works [38] thus improving the evaluation time to about four hours and the per-input amortized time to over seven seconds.

According to [32] in their study of efficiency of fully homomorphic encryption schemes, they improved on the previous works by introducing re-linearization and dimension-modulus reduction techniques. They argued that all somewhat homomorphic encryptions can be based on LWE by using the re-linearization technique since all previous schemes depended on the complexity assumptions based on ideals in various rings. They also deviated from the *squashing* approach used in previous works by introducing a new dimension-modulus reduction technique where the ciphertexts are shorten and thus reduces the decryption complexity of their scheme without bothering introducing any other assumptions.

III. METHODOLOGY, FINDINGS AND DISCUSSIONS

3.1 METHODOLOGY

In this paper, we focused on the time and space complexity of the algorithm, and the correctness of the solutions generated by the algorithm. Given that the objective was to develop an improved encryption scheme that enhances data security in cloud computing, an experimental research was adopted where combination of formal and theoretical

methodologies was found to be of much help. The newly developed scheme was coded and tested in a computing environment to ascertain that it is implementable. The data results generated was benchmarked with existing findings to prove its reliability and validity.

3.1.1 Test Lab Setup: The encryption algorithm of the new scheme was developed from the mathematical proofs realized. We considered both hardware and software of such test environment. On the hardware, it was tested on an AMD Quad-Core Processor 1.5 GHz with 4 GB DDR3 RAM Memory. The software component of the lab set up consisted of Windows 10 operating system (OS) and Java Runtime Environment for coding, compiling, debugging and test running of the code.

3.1.2 Metrics and Evaluation Parameters: The implementation of homomorphic encryption schemes considers several requirements and challenges. It is important to check the performance of the algorithm because of its impacts on the resource utilization. We evaluated the ACFHE Scheme using parameters: encryption time, decryption time, key generation time, and ciphertext size. This is in line with evaluation of an effective algorithm, that solves both time and space complexity.

3.1.3 Implementation Scenario: We focused on the use of integer values to demonstration the practicability of the new ACFHES algorithm. This was formed by the data representation in a classical computer, which operates on binary digits at the lowest definition.

3.1.4 Input Data Sets: With the scenario indicated above, we relied on integer data values to test the implementation of the ACFHE Scheme. This also supports the fact that the nature of data sets does not affect the performance of homomorphic encryption schemes but the size.

3.1.5 Benchmark Suite: The implementation of homomorphic encryption scheme can be benchmarked by other similar schemes that has been developed and implemented. ACFHES is asymmetrical probabilistic scheme that compares well with Gentry's construction, given that both are fully homomorphic encryption schemes.

3.2 ACFHES ALGORITHMN DEVELOPMENT

In the development of the ACFHES algorithm, the following steps were considered:

3.2.1 Designing the algorithm: We considered three main functions i.e. the key generation that relies on performing Euler Quotient function on prime numbers, the encryption function that concerns enciphering the input data sets, and the decryption function that decipher the ciphertext back into consumable output, plain data set.

3.2.2 Coding of the derived algorithm: This was done in Java programming language. Given that cloud servers are implemented web applications, Java as a programming language performs better. There were separate codes to capture the various component of the ACFHES scheme.

3.2.3 Benchmarking: It was compared against existing findings. An algorithm is required to meet the threshold of time and space complexity meaning the amount of computer processing power (CPU) and the amount of memory used in the run time.

3.2.4 Testing the code: The computer program generated from the developed algorithm was tested. The debugging process was considered to identify possible errors and appropriate correction done. The profiling or performance measurement was achieved by executing the program on data sets and measuring the time and space needed to realize results.

3.3 THE ACFHES ALGORITHM

This new scheme works in the polynomial ring $R = \mathbb{Z}[x]/(f(x))$ with $f(x) = \Phi_d(x)$, the d -th cyclotomic polynomial of degree $n = \phi(d)$. A plaintext is an element in the ring R_t for some modulus t where in this case t is taken as 2. A ciphertext in this scheme consists of elements in the ring R_q where q is the larger modulus. The security of this scheme is determined by the degree n of f , the size of q , and by the probability distributions. This ACFHE scheme uses an encryption scheme and two additional functions ADD and COMPO to perform function evaluations homomorphically on the encrypted data. The functions used in this scheme are summarized as below:

- ParamsGen (λ): for a given security parameter, choose a polynomial $\Phi_d(x)$, ciphertext modulus q and plaintext modulus t , and distribution key χ_{key} . Return the system parameters $(\Phi_d(x), q, t, \chi_{key})$ and set the plaintext modulus $t=2$.

- KeyGen ($\Phi d(x)$, q , t , χ_{key} , w): Sample polynomial s from χ_{key} , sample $a \leftarrow R_q$, uniformly at random. Compute $b = [-as]_q$. The public key consists of two polynomials $pk = \{b, a\}$ and the secret key $sk = s$.
- Encrypt (pk, m): the input message $m \in R_t$ is first encoded into polynomial $\Delta m \in R_t$ with $\Delta = \lfloor q/t \rfloor$. The ciphertext is the pair of polynomials $c = \{c_0, c_1\}$.
- Decrypt (sk, c): the polynomial is computed first $m = [c_0 + sc_1]_q$ and then when $t = 2$, recover the plaintext message m by a decoding the coefficients of m . this decoding operation checks if the coefficients is in $(q/4, 3q/4)$ for a 1 bit and 0 bit otherwise.
- Add (c_1, c_2): for two ciphertexts $c_0 = \{c_{0,0}, c_{1,0}\}$ and $c_1 = \{c_{1,0}, c_{1,1}\}$, return $c = \{c_{0,0} + c_{1,0}, c_{1,0} + c_{1,1}\}$.
- Compo (c_1, c_2): compute $c_{compo} = \{c_{0,1}, c_{1,1}\}$ where $c_0 = \lfloor t/q.c_{1,0}.c_{2,0} \rfloor/q$, $c_1 = \lfloor t/q.(c_{1,0}.c_{2,1} + c_{1,1}.c_{2,0}) \rfloor/q$, and $c_2 = \lfloor t/q.c_{1,1}.c_{2,1} \rfloor/q$.

In this scheme, only one bit is encrypted in one ciphertext and the encrypted bit remains in the least significant coefficient of the ciphertext polynomial. When it comes to decryption, only the coefficient of the least significant coefficient of m is decoded. The key generation algorithm generates the public and private keys that are to be used in the encryption process. This is necessary to enable achieve the key to be used in encryption and decryption process. The condition to be satisfied for the two chosen prime numbers is that both primes need to be of equivalent length, that is, $p, q \in 1 \parallel \{0, 1\}^{s-1}$ with s being the security parameter. Usually there are two ways of selecting the g generator. The other way is to use the equation $g = (\alpha n + 1)\beta^n \bmod n^2$. The public key is (n, g) and the private key is (λ, μ) . The encryption algorithm was the one used in the encipherment process, which provided the desired ciphertext c . The integer values m and r were provided and used in the calculation of ciphertext. The computed value was then displayed. The decryption algorithm above was used in the decipherment process, which reversed the ciphertext c to plaintext m . The calculated plaintext m was displayed to confirm the original text before the encryption process. In Addition algorithm, the homomorphic property of the algorithm is tested. The entered plaintext value is summed to prove the homomorphic property when compared. In Composition algorithm, the composition property is tested by computing the composition of the two values and this proves the homomorphic property of the algorithm. The values entered produce the composed results thus makes the plaintext “invisible” to anybody who access it.

3.4 IMPLEMENTATION AND TESTING OF ACFHES

The AFCHEs was benchmarked with other homomorphic encryption schemes for purposes of evaluating its efficiency. Efficiency is relative and not absolute, and can be explained under two main guidelines: time complexity and space complexity. The time complexity explains the time taken to complete the encryption and decryption process. The space complexity define the consumed computing storage resources at the encryption and decryption time. It is about the number of memory cells needed to carry out computational steps required to solve an instance of a problem excluding the space allocated to hold the inputs [39]. In this study, we adopted the simulational method since other homomorphic encryption schemes have been implemented and efficiency results are available therefore the remaining bit is to run results for ACFHES and compare them against existing ones. Moreover the computational power of implementation environment adopted here provided enough ground for effective comparison given that it was done under minimal specifications. The measurement of both encryption and decryption time can be achieved through the use of java application using smart card I/O API available in Java SE 6 and the System.nanoTime() method included in the API, which returns the current time in nanoseconds of the most precise available system time. The nanoTime which measures elapsed time and depends on the underlying architecture is significantly more accurate than currentTimeMillis that measures wall-clock time but it is an expensive call as well. Therefore System.nanoTime() guarantees a nanosecond precise time relative to arbitrary point. The developed code when run generated results as shown in table below. Each algorithm were timed differently to ascertain the computational time elapsed in each case. The data results were obtained after carrying out 20 times with an elapse time difference of 1 minute. This was done to enable “refreshing” the computational activity that may have been ignited by the previous operation. A mean value was computed to arrive at the average operation timing as captured in the table 4 below.

Table 4 Operation Timings in Milliseconds

Operation	$\lambda = 512$	$\lambda = 1024$	$\lambda = 2048$
KeyGen	15.83	93.07	2056.19
Encrypt	15.81	133.31	1453.77
Decrypt	1.45	1.87	2.36
Add	1.46	1.99	2.44
Compo	1.49	2.35	2.82

From the data displayed in Table 4 above, the algorithms for addition, composition and decryption are atomic operations for integer handling, and the most time consuming operations are the KeyGen and Encrypt since the

determination of a prime number takes relatively long runtime period. It took fairly longer to generate keys and encrypt the input data as compared to decryption time and time for testing the homomorphic properties: addition and composition.

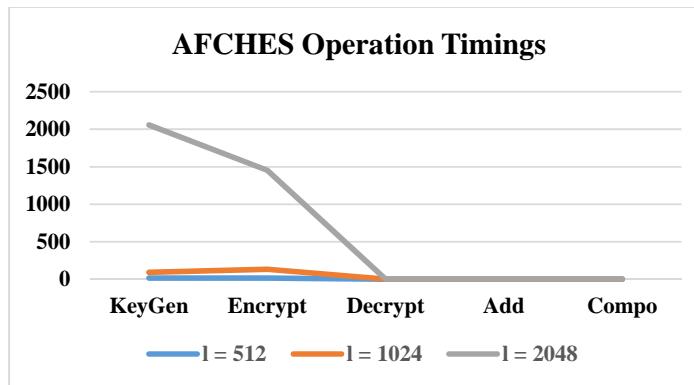


Fig. 7 AFCIES Operation Timings

Fig. 7 above was a graphical interpretation of the data captured in table 4 based on the considered basic operations of the scheme. The same data results were compared with a similar experiment done under Java 6 SE on a 2.4 GHz Core 2 Duo with 3 MB L2 cache and 4 GB RAM [40]. Their results were as also in milliseconds and were as follows:

Table 5 Timings for Basic Operations [55]

Operation	$\lambda = 512$	$\lambda = 1024$	$\lambda = 2048$
KeyGen	35	270	4242
Encrypt	35	301	3218
Decrypt	1	1	1
Add	1	1	1
Mult	1	1	1

The analysis of the same data from [40] was also graphical presented as in Fig. 8 below. The two graphs shows that the values obtained from AFCIES were better in terms of efficiency as measured compared to the earlier one done by [40].

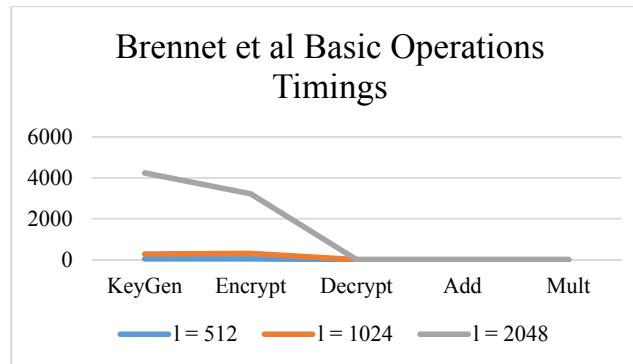


Fig 8 Basic Operations Timings [55]

Partial homomorphic operations are found to restrict the range of circuits that they support since they depend on one mathematical operations whereas fully homomorphic encryption schemes evaluate any Boolean circuits but computation speed and ciphertext size found to affects their efficiency. There are requirements that supports the effective evaluation of homomorphic schemes, that is, versatility, speed, and ciphertext size in an encryption scheme.

Table 6 Homomorphic Encryption Schemes Comparison

Type	Versatility	Speed	Ciphertext Size
PHE	Low	Fast	Small
FHE	High	Slow	Large
AFCIES	High	Faster	Medium

IV CONCLUSIONS AND FUTURE WORK

The study has reviewed both cloud computing and homomorphic encryption schemes, and shown how significant their relationship advantage data security. We have also appreciated previous contributions that has been made to improve security of data while in cloud computing. Our findings still agree with the fact that there is need for a better candidate for improving security of data in cloud computing and formed our main motivation. The main focus was to demonstrate how suitable the newly developed Addition-Composition Fully Homomorphic Scheme for securing data within cloud computing. The new scheme has posted impressive results, which when compared with previous ones, performs much better. It is also important to note that our testbed used a less superior computing environment thus point to high possibility of better results when both are tested on same environment. For future work, the study recommends that the performance of our scheme can be tested further on computing environment with superior specifications to enable the registration of better results.

REFERENCES

- [1] Rittinghouse John and Ransom James, “Cloud Computing: Implementation, Management and Security”, CRC Press, Taylor & Francis Group, 2010, pp 340.
- [2] Kant C. and Sharma Y, “Enhanced Security Architecture for Cloud Data Security”. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 3 Issue 5 ISSN: 2277128X, 2013, pp. 570 – 575.
- [3] Vaquero L.M, Rodero-Merino L., Caceres J. and Lindner M. “A break in the clouds: towards a cloud definition”, in: ACM SIGCOMM Computer Communication Review, 2008, p.50-55.
- [4] Mollah B. M., Islam K.R., and Islam S.S. “Next generation of computing through cloud computing technology”, in: 2012 25th IEEE Canadian Conference on Electrical Computer Engineering (CCECE), May 2012, p.1-6.
- [5] Ukil, A., Jaydip S., Bera, D., and Pal, A. “A Distributed Intrusion Detection System for Wireless Ad Hoc Networks”. In *Proceedings of the 16th IEEE International Conference on Networking (ICON'08)*, pp. 1-5, 2008, New Delhi, India.
- [6] Agudo I, Nunez D., Giammatteo G., Rizomiliotis P., and Lambrinoudakis C. “Cryptography goes to the Cloud. Secure and Trust Computing, Data management, and Applications”, Vol. 187, 2011 pp 190 – 197. Springer Berlin Heidelberg.
- [7] Gentry Craig, “A Fully Homomorphic Encryption Scheme” PhD Thesis in Computer Science, Department of Computer Science, Stanford University, California, USA. 2009
- [8] Weinman J. “Axiomatic Cloud Theory”. Communications, Media and Entertainment Industry for Hewlett-Packard, 2011.
- [9] Pranita P. K. and Ubale V. S. “Cloud Computing Security Issues and Challenges” International Refereed Journal of Engineering and Science (IRJES). 2013. Vol. 2 Issue 2 pp 10 – 13.
- [10] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinsky, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M (2009). “Above the Clouds: A Berkley View of Cloud Computing”. Technical Report No. UCB/EECS-2009-28, Department of Electrical Engineering and Computer Sciences, University of California at Berkley. February 10, 2009. Available on line at: <http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf> (Accessed on: November 2016).
- [11] Cloud Security Alliance (CSA)’s “Security Guidance for Critical Areas of Focus in Cloud Computing”. CSA, April 2009. Available Online at: <https://cloudsecurityalliance.org/csaguide.pdf> (Accessed on: November 2017).
- [12] Navneet S. P. and Rekha B. S. “Software as a Service (SaaS): Security issues and Solutions”. International Journal of Computational Engineering Research (IJCER). ISSN (e): 2250 – 3005, Vol. 04, Issue, 6, June – 2014.
- [13] Seccombe A., Hutton A., Meisel A., Windel A., Mohammed A. and Licciardi A. “Security guidance for critical areas of focus in cloud computing”, v2.1. Cloud Security Alliance, 2009, 25 p. pp 3121–3124.
- [14] Tiwari P. K. and Mishra B. “Cloud Computing Security Issues, Challenges and Solution”. International Journal of Emerging Technology and Advanced Engineering. Vol. 2 Issue 8 ISSN: 2250-2459, 2012.

- [15] Mehmet T. S. and Harmancı A. E. "Security Problems of Platform-as-a-Service (PaaS) Clouds and Practical Solutions to the Problems". 2012 31st International Symposium on Reliable Distributed Systems, IEEE Computer Society, pp 463 – 468.
- [16] Devi T. and Ganesan R. "Platform-as-a-Service (PaaS): Model and Security Issues". TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol. 15, No. 1, July 2015, pp. 151 ~ 161.
- [17] Gnanavelu D. and Gunasekaran G. "Survey on Security Issues and Solutions in Cloud Computing". International Journal of Computer Trends and Technology. 2014. Vol. 8 No. 3. Pp 126 – 130. ISSN: 2231-2803.
- [18] Jaydip S. and Sengupta, I. "Autonomous Agent-Based Distributed Fault-Tolerant Intrusion Detection System". In *Proceedings of the 2nd International Conference on Distributed Computing and Internet Technology (ICDCIT'05)*, pp. 125-131, December, 2005, Bhubaneswar, India. Springer LNCS Vol. 3186.
- [19] Jaydip S., Sengupta I., and Chowdhury P. R. "An Architecture of a Distributed Intrusion Detection System Using Cooperating Agents". In *Proceedings of the International Conference on Computing and Informatics (ICOCI'06)*, pp. 1-6, June, 2006, Kuala Lumpur, Malaysia.
- [20] Ukil A., Jaydip S., Bera D., and Pal A. "A Distributed Intrusion Detection System for Wireless Ad Hoc Networks". In *Proceedings of the 16th IEEE International Conference on Networking (ICON'08)*, pp. 1-5, December 2005, New Delhi, India.
- [21] Jaydip S. "An Agent-Based Intrusion Detection System for Local Area Networks". *International Journal of Communication Networks and Information Security (IJCNS)*, Vol 2, No 2, pp. 128-140, August 2010.
- [22] Jaydip S. "A Robust and Fault-Tolerant Distributed Intrusion Detection System". In *Proceedings of the 1st International Conference on Parallel, Distributed and Grid Computing (PDGC'10)*, pp. 123-128, October 2010, Waknaghat, India.
- [23] Trusted Computing Group (TCG)'s White Paper. "Cloud Computing and Security- A Natural Match". Available online at: <http://www.trustedcomputinggroup.org> (Accessed on; November 2017).
- [24] Jaydip S. "Security and Privacy Issues in Cloud Computing". Innovation Labs, Tata Consultancy Services Ltd, Kolkata, India. 2015.
- [25] Kumar K. S., Anjaneyulu N. and Venkanna. "Security Issues in Cloud Computing and study on Encryption Method". International Journal of Emerging Trends and Technology in Computer Science. 2013. Vol. 2 Issues 3, pp 442 – 446, ISSN: 2278 – 6856.
- [26] Shucheng Y., Cong W., Kui R. and Wenjing L. "Achieving Secure, Scalable and fine-grained data access control in cloud computing", in: IN-FOCOM, 2010 Proceedings IEEE, 2010.p.1-9.
- [27] Velumadhava R. R. and Selvamani K. "Data Security Challenges and Its Solutions in Cloud Computing". International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015). Procedia Computer Science 48, pp 204 – 209.
- [28] Chakraborty N. "Cloud Security using Homomorphic Encryption". National Conference on Advances in Computing, Networking and Security, 2013.
- [29] Rivest R. L., Adleman L. and Dertouzos M. L. "On Data Banks and Privacy Homomorphisms", chapter on Data Banks and Privacy Homomorphisms, 1978, pp. 169 – 180, Academic Press.
- [30] Boneh D., Goh E. and Nissim K. "Evaluating 2-DNF Formulas on Ciphertexts". In Theory of Cryptography Conference. 2005.
- [31] Lopez –Alt A., Tromer E., and Vaikuntanathan V. "On-the-Fly Multiparty Computation on the Cloud via Multikey Fully Homomorphic Encryption". In STOC 2012 (ACM).
- [32] Djik M., Gentry C., Halevi S., and Vaikuntanathan V. "Fully homomorphic encryption over the integers". In Proc. of Eurocrypt, Vol. 6110 of LNCS, 2010, pp 24-43. Springer.
- [33] Liu. "Practical Fully Homomorphic Encryption without Noise Reduction". IACR Cryptology ePrint Archive, vol. 468, 2015.
- [34] Jing Li and Licheng Wang. "Noiseless Fully Homomorphic Encryption". 2017. Cryptology ePrint Archive.

[35] Chungseng. "Attack on Fully Homomorphic Encryption Scheme over Integers". Cryptography and Security, 2012, pp 24.

[36] Chungseng. "Attack on Fully Homomorphic Encryption Scheme over Integers". Cryptography and Security, 2012, pp 24.

[37] Gentry C. and Halevi S. "Implementing Gentry's Fully-Homomorphic Encryption Scheme". EUROCRYPT'11 Proceedings of the 30th Annual International Conference on Theory and Applications of Cryptographic Techniques: Advances in Cryptology. ISBN: 978-3-642-20464-7. Springer-Verlog Berlin, Heidelberg. 2010.

[38] Dai W., Doroz Y, and Sunar B "Accelerating NTRU based Homomorphic Encryption using GPUs". Proceedings on 2014 IEEE High Performance Extreme Computing Conference, pp 1 – 6, ISBN: 978-1-4799-6232-7.

[39] Horowitz E., Sahni S., and Saguthevar R. "Fundamentals of Computer Algorithms". Galgotia Publications pvt Ltd, India.1998.

[40] Brenner M, Wiebelitz J, Voigt G, and Smith M. "Secret Program Execution in the Cloud Applying Homomorphic Encryption". 5th IEEE International Conference on Digital Ecosystems and Technologies, Daejeon, Korea. 2011. pp 114 – 119. ISBN: 978-1-4577-0872-5.

Conversion Prediction for Advertisement Recommendation using Expectation Maximization

Sejal D

Department of Computer Science and Engineering
B N M Institute of Technology,
Bangalore
Email: sej_nim@yahoo.co.in

Shradha G

Venugopal K R
Department of Computer Science and Engineering
University Visvesvaraya College of Engineering,
Bangalore University,
Bangalore-560001

S S Iyengar

Florida International University, USA

L M Patnaik

INSA Senior Scientist
National Institute of Advanced Studies
IISc Campus, Bangalore

Abstract—Advertiser has to understand the purchase requirement of the users who are looking for a particular service to recommend advertisement. Once the users' demand is identified, advertisers can target those users with appropriate query. In this paper, predicting conversion in advertising using expectation maximization [PCAEM] model is proposed to provide influence of their advertising campaigns to the advertisers by understanding hidden topics in search terms with respect to the time period. Query terms present in search log are used to construct vocabulary. Expectation Maximization technique is used to learn hidden topics from the vocabulary. Least Absolute Shrinkage and Selection Operator (LASSO) is used to predict total number of conversion. Experiment results show that PCAEM model outperforms TopicMachine model by reducing Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for prediction.

Keywords—Advertisement recommendation, Conversion Prediction, Expectation Maximization

I. INTRODUCTION

Internet provides information to users; it takes input as a query and gives related results. Many service providers use this characteristics of the Internet to their advantage. Consider a query *create an android app for generating recipe* is entered by a user to the search engine. The user may be looking for online tutorials or classes for learning android programming. It is assumed that such a user is willing to pay for a service that can help him to learn Android program. The service provider that provides such services wants his advertisement to appear as the most relevant result for this query.

The query given to a search engine is called as the search term and resulting advertisements are called as keyword based advertising. The search engine market (SEM) agencies connect the two sides of this same coin, i.e., the customers and the service providers. The SEM agencies maintain a large collection of keywords for their clients, the service providers. As the services provided by these clients grow, the number of keywords also increase and need to be managed. SEM agencies conduct experiments on the keywords to test their

weight, relevance and usage. These tests help the clients to select the best variant for the keywords related to the services they provide. However, with the increase in the number of keywords and many variants to consider, the advertisement management becomes a burden on even the most experienced advertisers.

The advertiser need to understand the underlying requirements of the users from the queries. These requirements are usually the descriptions of service. From the query, *create an android app for generating recipe*, the user might actually require *I want to create an android application where I can give picture of the ingredients that are available to me as an input and the generated application gives the relevant recipes out of these ingredients*. The result of the query must be relevant based on the information need present in it and not because it contains the keywords of the description.

Motivation: Advertisers have to analyze the purchase requirement of the targeted users, that helps them to aim those users with appropriate search terms. Advertisers analyze query logs, search keywords reports and trend reports to determine relevant search keywords. It is very difficult to understand the latent need of the user from a few words in the query and how an advertiser characterizes an information need [1], [2]. Hence, it is necessary to develop a model on top of the search campaigns that the advertisers can examine the consequence of the topical changes in trends over the time and target new market.

Contribution: In this paper, predicting conversion in advertising using expectation maximization [PCAEM] model is proposed to provide the effectiveness of their advertising campaigns to the advertisers with respect to time period. Vocabulary is constructed from the query keywords present in search query log. Gaussians are assumed for topic assignment and likelihood function is computed to find topic distribution of a query. Topic distribution of the Gaussians with respect to time period is defined with Topic Proportion Vector. Least

Absolute Shrinkage and Selection Operator (Lasso) is used to predict total number of conversion.

Organization: This paper is organized as follows: Various probabilistic topic modeling models are studied in section 2. Expectation Maximization (EM) and Latent Dirichlet Allocation (LDA) models are discussed in section 3. Prediction Conversion in Advertising using Expectation Maximization Model and Algorithm is presented in section 4. Section 5 discusses about experiment set-up, performance metrics and results analysis. Conclusions are presented in section 6.

II. RELATED WORKS

In this section, various probabilistic models for topic modelling are reviewed. Documents topics can be discovered by clustering the similar documents. A non-parametric clustering algorithm is proposed based on local shrinking in that the number of convergent points are used as number of clusters [3]. The idea is to transform data points towards denser regions. The Shrinking process is based on K-nearest neighbour method. Value of K is selected automatically based on optimized value of Silhouette and CH index. The non-parametric clustering algorithm out-performs traditional algorithms for all given data sets even when they are provided with true number of clusters as parameters. A clustering algorithm that can identify categories from the query keywords and assign advertisement to them is proposed [4]. This algorithm is based on Bernoulli distribution. Beta priors are used to maintain discrete probability distribution to assign clusters.

Document Influence Model (DIM) model is proposed based on Dynamic Topic Model (DTM) to find topics [5] over the time. Experiments are conducted on cited documents. This model computes the sequences of topics, posterior distribution of the latent variables and the per-document influence values. It shows how past articles decide the varying influence on future articles. Article's influence value is the hidden variable and the influential articles are identified by posterior inference.

Latent Dirichlet allocation (LDA) model is widely used to assign topics to the documents and it is an unsupervised model in which words in the documents are modelled. A supervised latent Dirichlet allocation (sLDA) model [6] is introduced that accepts various response types. The response variable can be movie rating, count of number of users who selected an article important, document category. Hidden topics are discovered by combining document and response and later used to predict the response variables. The sLDA model is limited to one variable association with document. A generative labelled LDA (L-LDA) model [7] is presented with multi-label supervision. Each label is associated with one topic directly. This model is a combination of supervised LDA and mixture model of Multinomial Naive Bayes. This L-LDA model is used to assign topics to Twitter profiles correctly and find similarity of profile pairs [8]. It also re-ranks Twitter blogs and suggest new users to follow. This L-LDA performs similar to Support Vector Machine and it outperforms when training data is limited.

Search Engine Marketing [SEM] agencies manage thousands of keywords for their clients. Advertisement brokers provided a management dashboard interface that allows them to change the search campaign attributes. Advertisers then use this dashboard to create test variants for various bid choices, keywords ideas etc. Controlled experiments are performed to reveal best performing variants. As the number of keywords increases, the test variants increases and the task of campaign management becomes burdensome. The advertisers need to understand the intent of the users in order to target them with particular services. Ahmet Bulut has proposed a framework called TopicMachine [9] which enables SEMs to scale and optimize for conversions. The TopicMachine uses LDA to reveal the hidden intent of the search terms that best matches client with its users and a Lasso-based predictor that predicts the conversion.

LDA models do not find the correlations between topics. Li et al. [10] introduced Pachinko Allocation Model (PAM) which captures the relation between topics by using directed acyclic graph (DAG). Each leaf in the DAG represents a word in a vocabulary and each internal node represents a relation between either words or topics. PAM with DAG does not represent the word's topical distribution that is present in multiple topics. Hierarchical PAM (hPAM) [11] arranges topics in hierarchies. This model combines the hierarchical nature of hLDA with the topic mixing abilities of PAM. In hPAM, each node is associated with distribution over the vocabulary. The resulting model is effective at discovering mixtures of topic hierarchies.

A hierarchical algorithm which represents documents as a hierarchy of latent topics computed with Dirichlet process [12]. It is based on Bayesian priors and hierarchy of topics is derived without estimating depth of the hierarchy and branching at each level. The internal nodes represents words and topics probability distribution and vocabulary clustering is performed. Leaf nodes represents words distribution in a corpus hierarchical topic clustering. This model does not restrict on topic usage, allows multiple inheritance between topics and internal nodes are modelled as subtopics and words distribution. Method proposed in [13], [14], [15] can be used for prediction.

Ravi. et al. [16] proposed a probabilistic method that generates bid phrases for online advertising. This model first trained on search query log and generates well-formed bid phrases. Next, it generates novel bid phrases from webpages and corpus of bid phrases. Fujita et al. [17] proposed a method that generates shop-specific ad listing by incorporating promotional text data for restaurant domain. Experiments result shows that the click through rate is higher for automatically generated ad than template based ad.

III. BACKGROUND

In this section, Expectation Maximization (EM) that is used to construct the proposed model and Latent Dirichlet Allocation (LDA) which is used for comparison with the

proposed model are explained.

A. Expectation Maximization (EM)

Expectation maximization is an iterative process used to compute the maximum likelihood of parameters in a statistical model, especially in models where some data is unobserved [18], [19], [20]. It is used as a clustering technique. Each cluster can be mathematically represented as a probability distribution, characterized by its mean and variance. EM assumes that the data is generated by a mixture of underlying probability distribution. It assigns each object to a cluster depending on the likelihood of its membership. There are no restrictive boundaries between the clusters, i.e., an object can belong to multiple clusters with same or different probability of membership.

EM algorithm comprises of two distinct steps; the Expectation step (*E – Step*) and the Maximization step (*M – Step*). In the *E – step*, the estimated parameters are used to compute the likelihood of the model. In the *M – step*, the likelihood computed in the previous step are used to determine new values for the model parameters. The algorithm converges when there is no significant change in the model.

Consider two coins *A* and *B* are tossed *five* times, the probability *p* of head coming up on these tosses is known as $P=p_1, p_2...p_5$. The goal is to estimate the identity of the coin for these tosses as $Q=q_1, q_2...q_5$ where q_i is the identity of the coin for the *i*th toss. Here *Q* is the latent or hidden variable.

For this problem, computing the proportion of heads for each coin is not a viable option. However, the parameter estimation with incomplete data can be transformed to maximum likelihood estimation with complete data using EM. The following steps are involved:

- 1) Initial values for P_A and P_B are selected randomly.
- 2) For each of the five flips, estimate the likelihood of the flip being made using coin *A* or *B*.
- 3) Assuming these completions to be correct, calculate the new values for parameters P_A and P_B .
- 4) Repeat step 2 and 3 until the difference between the previous and the current estimates is negligible. This defines the convergence point of the problem.

B. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [21] is a common method for topic modelling. It is a generative model which separates words into different topics from a *corpus*. LDA assumes that each document is a mixture of different topics, where each topic has some particular probability of generating a particular word. A document is assumed to be a *bag of words* pulled from a distribution selected by a Dirichlet process.

Consider a corpus *C* which is a collection of documents *D*. The goal is to discover *K* topics from *C*, topic distribution

of *D* and words associated with each topic. LDA performs the following step to achieve this goal:

- 1) Each word *W* in a document *D* is randomly assigned to a topic *T*.
- 2) This random assignment gives a basic structure of the goal, i.e., initial topic distribution of the document and the word distribution for the topics.
- 3) For each word *W* in *D*, the conditional probabilities $P(T/D)$ and $P(W/T)$ are estimated. $P(T/D)$ represents the probability of the words in *D* which are assigned to topic *T*. $P(W/T)$ is the probability of *W* assigned to topic *T* over all documents in the corpus.
- 4) Each word is re-sampled and it is assumed that the topic assignment for the current word is incorrect. Word *W* is then assigned a new topic by computing $P(T/D)*p(W/T)$ that gives the probability of topic *T* containing word *W*.
- 5) The previous step is iterated until a steady state is reached. The assignments at this state are used to estimate the topic distribution of the documents and the word distribution of the topics.

IV. PREDICTION CONVERSION IN ADVERTISING USING EXPECTATION MAXIMIZATION MODEL AND ALGORITHM

A. Problem Definition

Given a search log *l* and number of epoch *e*, the objective is to predict the total conversion in advertising.

B. Prediction Conversion in Advertising using Expectation Maximization Model

Predicting conversion in advertising using expectation maximization (PCAEM) model provides the advertisers with information on the effectiveness of their advertising campaigns. This effectiveness can be measured by estimating the number of conversions with respect to time period. This framework helps the advertisers in making business decisions. The PCAEM model has the following steps:

Step 1: Building a Vocabulary

In this step, the queries from the query log are considered to build a vocabulary. Stopwords *a, an, the, for etc.* are used as grammar constructs and don't convey any meaning to the query. Hence, they are removed from the queries to obtain the words that convey the requirements of the users. The vocabulary *V* is built from the remaining words in the queries after removal of stopwords and the frequency of occurrence of those words are computed. Consider a query *create an android app for generating recipe*, in which the words *an* and *for* are removed. After pre-processing, the query now becomes *create android app generating recipe*.

Step 2 : Gaussian Assumption for Defining Topics

A query is a mixture of various topics. Hence it is necessary to determine the latent topics in a query. Consider the query *create an android app for generating recipe*, it comprises of words related to both *technology* as well as *food*. In EM, topics are represented as Gaussians $Gauss[K]$, where *K* is the number of topics. A Gaussian is characterized by

its mean and variance. Gaussians contain all the words in the vocabulary, in which each word is randomly assigned a topic between 1 to K . The initial probabilities of the words belonging to topic i in Gaussians are equal, where $i \in 1,..k$. The initial mean of a Gaussian and covariance between the Gaussians is computed as shown in Function 1.

Function 1: Initial Mean and Covariance Generation

Function: InitialTopic

Data: Consider Vocabulary V and Number of Topics K . Initial Mean and Covariance Matrix are generated.

Let n = number of Gaussians = number of topics
 $mean[n][K]$ = mean of each topic in K for n Gaussian
 $covariance[n][n]$ = covariance between Gaussians
 $Gauss[n]$ = n Gaussians
 V_{size} = vocabulary size
for $i = 1$ to n **do**
 $Gauss[i]$ = All vocabulary words
 for $j = 1$ to V_{size} **do**
 $Gauss[i].word[j]$ = random topic assignment
 between 1 to K
 for $L = 1$ to K **do**
 Let Num_{word} = Number of words in Gaussian i with topic K
 Let Tot_{word} = Total number of words in Gaussian i
 Calculate mean for topic L in $Gauss[i]$ =
 $Mean[i][L] = \frac{Num_{word}}{Tot_{word}}$
for $i = 1$ to n **do**
 for $j = 1$ to n **do**
 Calculate covariance($Gauss[i], Gauss[j]$) using
 Equation 1

Here, $Prob_{Gauss_i}$ is the probability of $word_c$ in $Gauss[i]$, $Prob_{Gauss_j}$ is the probability of $word_c$ in $Gauss[j]$, $topic(word_c, i)$ is the topic $word_c$ in Gaussian i and $mean[i][topic(word_c, i)]$ is the mean of $topic(word_c, i)$ in Gaussian i .

Step 3 : Compute Likelihood to find Topic Distribution of a Query

In the previous step, the topics defined in the Gaussians are estimated. Using these Gaussians, the topic distribution of the queries is computed. All the queries are assumed to be stationary points in a space. This space also contains the Gaussians with equal probability $P(c)$. As the Gaussians move around and take shape, the conditional probability of occurrence $P(qi/c)$ of a query qi in the Gaussian c changes. A query qi can belong to more than one topic. Consider the query *create an android app for generating recipe*, the words *android* and *app* come from the *technology* topic and the word *recipe* comes from the *food* topic, hence this query

$$covariance_c(j, k) = \sum_{j=1}^{Q_{num}} \left(\frac{PCQ(c, i)}{nP(c)} \right)^* \\ (q_{i,j} - mean(c, j)) * (q_{j,k} - mean(c, k)) \quad (6)$$

occurs in the *technology* topic with probability (2/5) and in the *food* topic with probability (1/5). The likelihood of a query occurring in a topic is computed as shown in Function 2.

Function 2: Query Topic Likelihood

Function: QueryTopicLikelihood

Data: Consider Mean matrix $mean[n][k]$, Covariance matrix $cov[n][n]$, Number of queries Q_{num} and n is the number of Gaussians. Conditional Probability of Topic C for given Query q PCQ , Conditional Probability of Query q for given Topic C PQC and Probability of Gaussian $P(C)$ are generated.

```

for  $i = 1$  to  $Q_{num}$  do
    for  $c = 1$  to  $K$  do
         $\downarrow$  Compute  $PQC(i, c)$  using Equation 2
for  $c = 1$  to  $K$  do
    for  $i = 1$  to  $Q_{num}$  do
         $\downarrow$  Compute  $PCQ(c, i)$  using Equation 3
for  $c = 1$  to  $n$  do
     $\downarrow$  Compute  $P(c)$  using Equation 4

```

Here, $| cov_c |$ is the determinant of $covariance_c$.

$$PCQ(c, i) = \frac{PQC(i, c) * P(c)}{\sum_{c'=1}^n PQC(i, c') * P(c')} \quad (3)$$

$$P(c) = \frac{1}{n} \sum_{j=1}^{Q_{num}} PCQ(c, j) \quad (4)$$

Step 4 : Compute New Definition of Topics

The initial mean and covariance of the Gaussians give a basic structure to the topic. The conditional probabilities estimated in the previous step influence the mean and covariance of the Gaussians such that the mean of the topics in the Gaussian changes. Hence, the Gaussians need to be recomputed with respect to the conditional probabilities. The mean and covariance of the Gaussians is computed using Equation 5 and 6.

$$mean(c, j) = \sum_{i=1}^{Q_{num}} \left[\frac{PCQ(c, i)}{cP(c)} \right] q_{i,j} \quad (5)$$

Step 5 : Test for Convergence of the Model

Convergence signifies a stable state of the model. After certain number of iterations the Gaussians gain a definite shape and position in space where the change in the conditional

$$covariance_c(i, j) = [(Prob_{Gauss_i}) - (mean[i][topic(word_c, i))]] * [(Prob_{Gauss_j}) - (mean[j][topic(word_c, j))]] \quad (1)$$

$$PQC(i, c) = \frac{1}{\sqrt{2\pi | cov_c |}} \left\{ \frac{-1}{2} \left[\sum_{K=1}^n \sum_{j=1}^n ((q_{i,j} - mean(c, j)) * (q_{i,k} - mean(c, k))) * (cov_c)^{-1}_{j,k} \right] \right\} \quad (2)$$

probabilities is negligible. Hence, step 3 and 4 need to be iterated until the model converges.

Step 6 : Compute Topic Proportion Vector

Topic proportion vector TPV represents the topic distribution of the Gaussians in the model. It is obtained from the mean computed at the convergence. As this model predicts the total number of conversions for advertisement campaign in a particular time period which is called an *epoch*. TPV for each epoch is defined as $TPV[e][i]$, where e and i represent epoch and topic respectively. TPV for an epoch represents the weight of the topics in that epoch.

Step 7 : Conversion Prediction

A conversion occurs when a user clicks on the advertisement and performs some action that gives benefit to the advertiser. This type of user click is defined as Cost Per Click (CPC). In order to predict the total number of conversions, the model is trained using the *L₁ prior* also known as *Lasso* (Least Absolute Shrinkage and Selection Operator) [22]. The conversion for each epoch is estimated using Equation 7.

$$conversion = \frac{1}{2*T} * \| CPC - TPV * R \|_2^2 + \alpha * \| R \|_1 \quad (7)$$

Here, α is the hyper parameter, R is the Regression and CPC is the cost per click.

C. Algorithm

Predicting conversion in advertising using expectation maximization (PCAEM) algorithm as shown in Algorithm 1 has two phases : Pre-processing and Prediction. Pre-processing involves building vocabulary, Gaussian assumption, computing likelihood for topic distribution of queries, re-estimation of the Gaussians and constructing topic proposition matrix. Prediction involves Lasso based conversion prediction.

V. EXPERIMENTS

A. Data Collection

In this experiment, the same dataset as [9] is used. The data is collected from U.S ad U.K market from over a 34 week period. It consists of various queries ranging over multiple topics. This dataset consists of following information: Max. CPC, Keyword, Average position, Average CPC, Clicks, CTR, Cost and Impressions. Here, Click Through Rate (CTR) is the ratio of how often user clicks an advertisement that appears as a relevant result for the query. It evaluates the competence of

Algorithm 1: Predicting Conversion in Advertising using Expectation Maximization

```

Input : Query log  $l$ , Number of epoch  $N_e$ 
Output: Conversion prediction  $cp$ 

begin
    Pre-processing :
        Build vocabulary  $V$  by considering queries from log  $l$ 
    for  $m = 1$  to  $N_e$  do
        Assume  $K$  Gaussians to represent topics as explained in step 2.
        while model not converged do
            Estimate the likelihood to find topic distribution of each query as per step 3.
            Re-estimate Gaussians to find new definition of the topics as per step 4.
        for  $n = 1$  to  $K$  do
            Construct Topic Proportion Vector  $TPV[m][n]$  using the final Gaussian estimation after convergence.
    Prediction :
        Predict conversions using LASSO method as per step 7

```

the keywords and the advertisement. An impression signifies the relevance of an advertisement and is incremented every time it appears as a result for a searched query. There were four different campaigns with 13,898 unique search terms that resulted in 432 conversions in total; 29,821 clicks were received out of a total of 2,382,317 ad impressions. In the proposed model, keywords are used to generate the Topic Proportion Matrix and CPC influence the conversion. Hence, Keyword and Average CPC is used in this experiment.

B. Experiment Setup

The proposed model PCAEM uses probabilistic estimations for topic modelling. LDA also works with same principle, hence PCAEM is compared with TopicMachine [9].

The setup of PCAEM is as follows: Vocabulary is constructed from the keywords. The words in the vocabulary are ranked based on their frequency of occurrence. Consider the query *create android application for food suggestions*; this query gives outcomes related to existing applications that make food suggestions rather than tutorials to learn android programming. In this example, the word *create* is a

low frequency word. Thus, it can be concluded that words with low frequencies do not contribute significantly the topic proportion of a query thereby influencing the outcome. Hence, the size of the vocabulary is restricted to 500. Experiments are conducted with varying topic number $K = 5, 6, 7, 8, 9, 10, 11$. The initial probability of each topic is considered equal. Using these assumptions the topic distribution of each query is estimated. The model is iterated 10 times for convergence. It is observed that PCAEM model converges at the fifth iteration. Hence, the number of conversions are estimated after the fifth iteration. The data is collected for 34 weeks, hence the number of epochs are 34 as 1 epoch is for 1 week. The experiment is conducted by varying epoch size to 15, 20, 25, 30 and 34. Topic proportion vector is constructed for each epoch and used for conversion prediction.

The setup for TopicMachine is as follows: Vocabulary is constructed in a similar way as PCAEM. Initially, words in the queries are randomly assigned to topics. These assignments are used to approximate hidden variables using variational approximation iteratively. Variation threshold is set to define a convergence point for the model. Experiments are conducted by varying the threshold parameter to 0.1, 0.01 and 0.001. It is observed that definite topics are obtained when threshold is set to 0.001. Topic proportion vector is constructed for each epoch and used for conversion prediction.

C. Performance Metrics

The performance metrics used for comparison are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Predictive R^2 i.e., the coefficient of determination. The actual value for each epoch corresponds to the average CPC collected from the dataset. Consider E is the total number of epochs, dataset is partitioned into training set T and testing set T_e , then $offset = E * T_e / 100$. The predicted and actual value is denoted by p and a respectively.

RMSE is the root of the square of the difference between the values predicted by the model and the actual values as shown in Equation 8. MAE is the average of the absolute value of the difference between the values predicted by the model and the actual values as shown in Equation 9. Coefficient of determination is computed as shown in Equation 10. It measures the correctness of the model based on the proportion of deviation of predicted value from actual value.

$$RMSE = \sqrt{\sum_{i=E-offset+1}^E (a_i - p_i)^2} \quad (8)$$

$$MAE = \frac{1}{N} \sum_{i=E-offset+1}^E |p_i - a_i| \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=E-offset+1}^E (a_i - p_i)^2}{\sum_{i=E-offset+1}^E (a_i - mean(p_i))^2} \quad (10)$$

D. Performance Evaluation

In this section, experiment results are presented and discussed. Performance metrics are used to compare the results of the proposed model PCAEM and TopicMachine [9]. Experiments are conducted by varying training dataset to 70%, 75%, 80%, 85%, 90% and testing dataset to 30%, 25%, 20%, 15%, 10%. Experiments have been conducted on 4GB memory and Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz processor. The reproduction of TopicMachine does not get the same result as in [9], this discrepancy is mainly due to the system properties of the machine used and the programming language; all development is done in Java. Dataset used in the experiments for PCAEM and TopicMachine are the same as discussed in data collection. The performance is evaluated by computing the average of 20 independent runs.

RMSE, MAE and R^2 by Varying Number of Topic K

In order to measure the effect of the number of topics K , all the performance metrics are computed to measure the performance of PCAEM and TopicMachine. The value of K varies from 5 to 11. Table I shows the comparison of RMSE, MAE and R^2 values for both the methods by setting vocabulary size to 500, α to 0.01, epoch to 34 and training dataset to 85%. It is observed from the table that the number of topics does not affect much on the performance of PCAEM and TopicMachine. It is also observed from the table that the RMSE and MAE values are lesser in PCAEM than TopicMachine, hence the performance of PCAEM is better than TopicMachine.

RMSE, MAE and R^2 by Varying Training Dataset

In order to analyze the the performance of models with available training dataset, RMSE, MAE and R^2 is computed by increasing training dataset. Table II shows the comparison of RMSE, MAE and R^2 values for both the methods by setting vocabulary size to 500, α to 0.01, Topic number K to 10 and epoch to 34. It is observed from the Table II that as the training dataset increases the RMSE decreases and MAE increases marginally.

RMSE, MAE and R^2 by Varying epoch Size

In order to study the consequence of the size of the dataset on the performance of PCAEM and TopicMachine, RMSE, MAE and R^2 is computed by varying the epoch size. Table III shows the comparison of RMSE, MAE and R^2 values for both the methods by setting vocabulary size to 500, α to 0.01, Topic number K to 10 and training dataset to 85%. There is not much difference observed by varying size of the dataset.

PCAEM and TopicMachine Sensitivity to Hyper Parameter α

The hyper parameter α is used for prediction conversion, it is necessary to study the effect of changes in α affect the quality of model. RMSE, MAE and R^2 is computed by varying the α value for both the models. It is observed from Table II that RMSE values decreases when training dataset increases, hence, experiments are conducted by setting training dataset 90% and 85%. Table IV shows the comparison of RMSE, MAE and R^2 values for both the

TABLE I: RMSE, MAE and R^2 by Varying Number of Topic K for PCAEM and TopicMachine. The vocabulary size, α , epoch and training dataset are set to 500, 0.01, 34 and 85% respectively

Topic	PCAEM Method			TopicMachine		
	RMSE	MAE	R^2	RMSE	MAE	R^2
5	8.6142	4.3039	-18.6755	8.6898	4.3416	-19.0221
6	8.6071	4.3004	-18.6431	8.6820	4.3377	-18.9863
7	8.6054	4.2995	-18.6354	8.6839	4.3387	-18.9949
8	8.6040	4.2988	-18.6286	8.6838	4.3386	-18.9948
9	8.6034	4.2985	-18.626	8.6794	4.3364	-18.9743
10	8.6019	4.2978	-18.6193	8.6795	4.3365	-18.9751
11	8.6015	4.2975	-18.6172	8.6706	4.3321	-18.934

TABLE II: RMSE, MAE and R^2 by Varying Training dataset for PCAEM and TopicMachine. The vocabulary size, α , Topic Number K and epoch are set to 500, 0.01, 10 and 34 respectively

Training Dataset	PCAEM Method			TopicMachine		
	RMSE	MAE	R^2	RMSE	MAE	R^2
70	12.2351	4.0718	-6.1212	12.3420	4.1073	-6.2462
75	10.9441	4.1300	-7.7055	11.0403	4.1662	-7.8592
80	9.4557	4.2235	-11.8799	9.5403	4.2612	-12.1114
85	8.6019	4.2978	-18.6193	8.6795	4.3365	-18.9751
90	6.2202	4.3970	-40.7745	6.2770	4.4370	-41.5397

TABLE III: RMSE, MAE and R^2 by Varying epoch Size for PCAEM and TopicMachine. The vocabulary size, α , Topic Number K and training dataset are set to 500, 0.01, 10 and 85% respectively

epoch	PCAEM Method			TopicMachine		
	RMSE	MAE	R^2	RMSE	MAE	R^2
15	3.4425	3.4425	0.8266	3.5394	3.5394	0.8168
20	6.7551	4.7729	-19.9357	6.8263	4.8230	-20.379
25	5.0849	3.5898	-0.4554	5.1259	3.6187	-0.4790
30	6.7661	3.9063	-3.0550	6.6234	3.9394	-3.1240
34	8.6019	4.2978	-18.6193	8.6795	4.3365	-18.9751

methods by setting vocabulary size to 500, Topic number K to 10, epoch to 34 and training dataset to 90%. Table V shows the comparison of RMSE, MAE and R^2 values for both the methods by setting vocabulary size to 500, Topic number K to 10, epoch to 34 and training dataset to 85%. It is observed from the Table IV and V that PCAEM model is sensitive to hyper parameter α , but TopicMachine is not. PCAEM model is giving best result when α is set to 10.

It is observed from Table IV and V that PCAEM model is giving best results when α is set to 10. Hence, RMSE and MAE is computed again by varying training dataset, by varying the number of the topic K and by varying the epoch size by setting α value to 10. Figures 1 and 2 show the comparison of RMSE and MAE by Varying Training Dataset when α is set to 10 respectively. It is observed that the average value of RMSE is 4.4572 and 9.5752 of PCAEM and TopicMachine respectively. The average value of MAE is 1.9346 and 4.2614 of PCAEM and TopicMachine respectively.

Figures 3 and 4 show the comparison of RMSE and MAE by the varying number of topics K when α is set

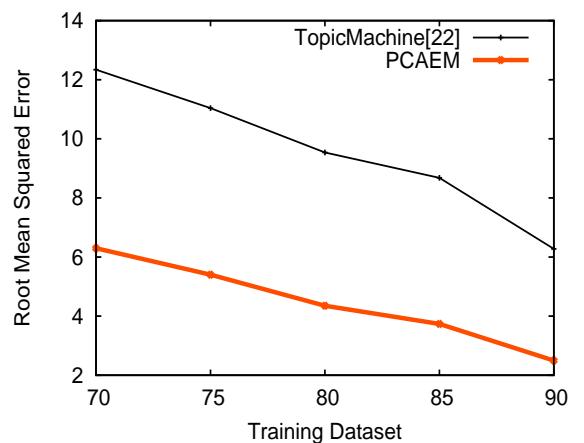


Fig. 1: RMSE by Varying Training Dataset When $\alpha = 10$

to 10 respectively. It is observed that the average value of RMSE is 3.2013 and 8.6815 of PCAEM and TopicMachine respectively. The average value of MAE is 1.5861 and 4.3375 of PCAEM and TopicMachine respectively.

TABLE IV: RMSE, MAE and R^2 by Varying Hyper Parameter α for PCAEM and TopicMachine. The vocabulary size, Topic Number K, epoch and training dataset are set to 500, 10, 34 and 90% respectively

α	PCAEM Method			TopicMachine		
	RMSE	MAE	R^2	RMSE	MAE	R^2
100	80.9580	57.2459	-7075.36	6.2769	4.4370	-41.5388
50	37.3657	26.4213	-1506.43	6.2778	4.4377	-41.5518
30	19.9290	14.0915	-427.807	6.2771	4.4371	-41.541
20	11.2110	7.9266	-134.7	6.2771	4.4372	-41.542
10	2.4962	1.7617	-5.7279	6.2773	4.4373	-41.5442
1	5.3574	3.7867	-29.9892	6.2772	4.4372	-41.5429
0.1	6.1418	4.3415	-39.7275	6.2778	4.4376	-41.5508
0.01	6.2202	4.3970	-40.7745	6.2778	4.4376	-41.551
0.001	6.2281	4.4025	-40.8799	6.2774	4.4374	-41.5461

TABLE V: RMSE, MAE and R^2 by Varying Hyper Parameter α for PCAEM and TopicMachine. The vocabulary size, Topic Number K, epoch and training dataset are set to 500, 10, 34 and 85% respectively

α	PCAEM Method			TopicMachine		
	RMSE	MAE	R^2	RMSE	MAE	R^2
100	114.6907	57.3451	-3486.75	8.6793	4.3364	-18.974
50	53.0421	26.5205	-744.988	8.6791	4.3362	-18.9728
30	28.3834	14.1907	-212.609	8.6790	4.3362	-18.9723
20	16.0551	8.0258	-67.3464	8.6797	4.3366	-18.976
10	3.7365	0.8609	-2.7018	8.6794	4.3364	-18.9746
1	7.3823	3.6874	-13.4503	8.6794	4.3365	-18.9746
0.1	8.4910	4.2423	-18.1167	8.6797	4.3366	-18.9758
0.01	8.6019	4.2978	-18.6193	8.6793	4.3364	-18.9738
0.001	8.6130	4.3033	-18.6699	8.6796	4.3365	-18.9755

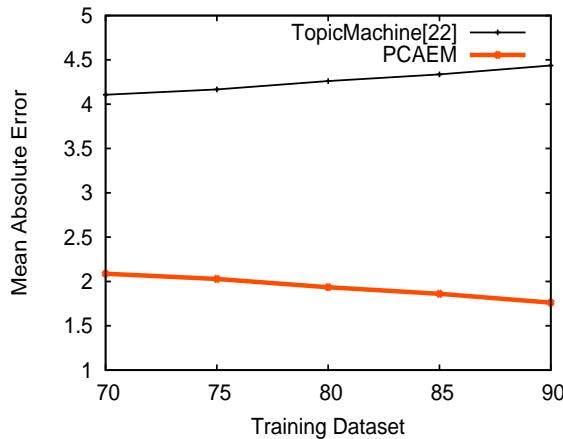


Fig. 2: MAE by Varying Training Dataset When $\alpha = 10$

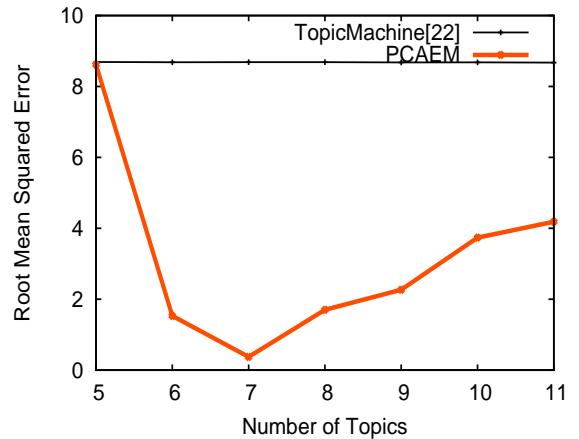


Fig. 3: RMSE by Varying Number of Topics When $\alpha = 10$

VI. CONCLUSIONS

Figures 5 and 6 show the comparison of RMSE and MAE by varying the *epoch* size when α is set to 10 respectively. It is observed from that the average value of RMSE 3.1952 and 6.1987 of PCAEM and TopicMachine respectively. The average value of MAE is 2.1568 and 4.0513 of PCAEM and TopicMachine respectively.

In this work, we present predicting conversion in advertising using expectation maximization [PCAEM] model to understand advertising campaigns' effectiveness to the advertisements over the time. Search query log is used to build vocabulary. Expectation Maximization method is used to find the hidden topics and topic distribution on search terms. Least Absolute Shrinkage and Selection Operator (LASSO) is used predict total number of conversion. Experiments are performed on

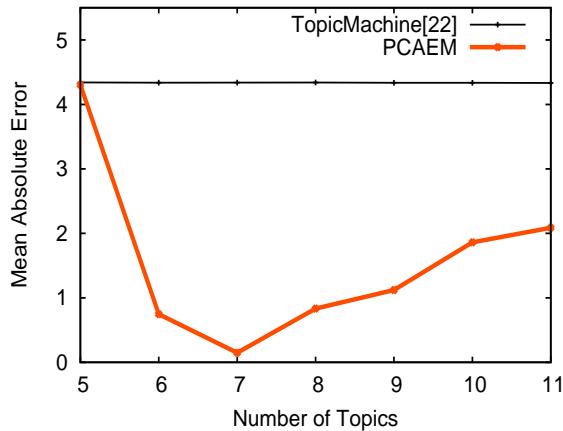


Fig. 4: MAE by Varying Number of Topics When $\alpha = 10$

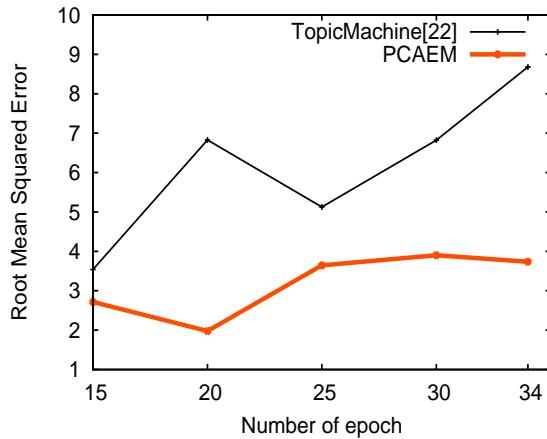


Fig. 5: RMSE by Varying epoch When $\alpha = 10$

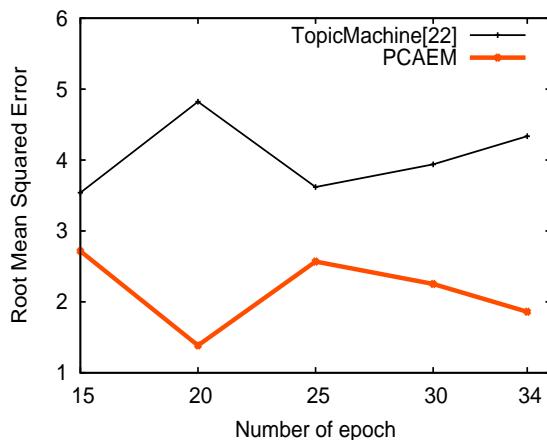


Fig. 6: MAE by Varying epoch When $\alpha = 10$

query data used in [9] which is collected from the U.K and the U.S market over 34 weeks. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Predictive R^2 are used as performance metrics. Experiment results are compared with TopicMachine [9]. The proposed method outperforms TopicMachine [9] by reducing RMSE and MAE. The PCAEM model is sensitive to hyper parameter used for prediction conversion while TopicMachine is not.

REFERENCES

- [1] B. Ribeiro-Neto, M. Cristo, P. B. Golher, and E. Silva de Moura, "Impedance Coupling in Content-targeted Advertising," *In the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 496–503, 2005.
- [2] P. Rusmevichientong and D. P. Williamson, "An Adaptive Algorithm for Selecting Profitable Keywords for Search-based Advertising Services," *Proceedings of the 7th ACM Conference on Electronic Commerce*, pp. 260–269, 2006.
- [3] X. Wang, W. Qiu, and R. H. Zamar, "An Iterative Non-parametric Clustering Algorithm based on Local Shrinking," *Computational Statistics and Data Analysis*, vol. 52, pp. 286–298, 2007.
- [4] A. Schwaighofer, J. Q. Candela, T. Borchert, T. Graepel, and R. Herbrich, "Scalable Clustering and Keyword Suggestion for Online Advertisements," *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, pp. 27–36, 2009.
- [5] S. Gerrish and D. M. Blei, "A Language-based Approach to Measuring Scholarly Impact," *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 375–382, 2010.
- [6] J. D. McAuliffe and D. M. Blei, "Supervised Topic Models," *Advances in Neural Information Processing Systems*, pp. 121–128, 2008.
- [7] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora," pp. 248–256, 2009.
- [8] D. Quercia, H. Askham, and J. Crowcroft, "TweetLDA: Supervised Topic Classification and Link Prediction in Twitter," in *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012, pp. 247–250.
- [9] A. Bulut, "TopicMachine: Conversion Prediction in Search Advertising using Latent Topic Models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 11, pp. 2846–2858, 2014.
- [10] W. Li and A. McCallum, "Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations," *Proceedings of the 23rd International Conference on Machine Learning*, pp. 577–584, 2006.
- [11] D. Mimmo, W. Li, and A. McCallum, "Mixtures of Hierarchical Topics with Pachinko Allocation," *Proceedings of the 24th International Conference on Machine Learning*, pp. 633–640, 2007.
- [12] E. Zavitsanos, G. Paliouras, and G. A. Vouros, "Non-parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2749–2775, 2011.
- [13] K. Srinivasa, K. Venugopal, and L. M. Patnaik, "An Efficient Fuzzy based Neuro-genetic Algorithm for Stock Market Prediction," *International Journal of Hybrid Intelligent Systems*, vol. 3, no. 2, pp. 63–81, 2006.
- [14] K. Srikanthiah, M. Roopa, N. K. Kumar, K. Venugopal, and L. Patnaik, "Automatic discovery and ranking of synonyms for search keywords in the web," *International Journal of Web Science*, vol. 2, no. 4, pp. 218–236, 2014.
- [15] K. Srikanthiah, N. K. Kumar, K. Venugopal, and L. M. Patnaik, "Web Caching and Prefetching with Cyclic Model Analysis of Web Object Sequences," *International Journal of Knowledge and Web Intelligence* 2, vol. 5, no. 1, pp. 76–103, 2014.
- [16] S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, and B. Pang, "Automatic Generation of Bid Phrases for Online Advertising," *Proceedings of the Third ACM International Conference on Web Search and Data mining*, pp. 341–350, 2010.

- [17] A. Fujita, K. Ikushima, S. Sato, R. Kamite, K. Ishiyama, and O. Tamachi, "Automatic Generation of Listing Ads by Reusing Promotional Texts," *Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business*, pp. 179–188, 2010.
- [18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
- [19] J. A. Bilmes *et al.*, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," *Journal on International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.
- [20] T. K. Moon, "The Expectation-Maximization Algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.
- [21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [22] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

An Android Application Studhelper for Engineering Students

Ishani Mukherjee¹, Aman Bansal², Mokshada Patra³, Rahul Pal⁴, Md.Khaja Mohiddin⁵

UG Scholars^{1,2,3,4}, Senior Assistant Professor⁵

Department of Electronics & Telecommunication Engineering

Bhilai Institute of Technology, Raipur

Raipur, Chhattisgarh, India

ishani.mukherjee21@gmail.com, mnbansal69@gmail.com, mpatra126@gmail.com, rprahulpal31@gmail.com
khwaja7388@gmail.com

Abstract—The paper has been presented with a system that is created on the android platform targeting the students studying in an engineering institute. The application is created for effortless day to day official work in an institute. Students will be served with the benefits like compilation of branch wise question papers, general aptitude questions, video lectures, newspapers, with some interesting features like parent-teacher portal, feedback system and ask your queries block. The application is designed using core java coding, layout is fabricated with xml extension, complete creation is done on android studio. Login authentication is developed via Firebase Auth, newspapers have been linked through their URL, video lectures are engrossed with API and a separate website is created for parent-teacher interaction. Based on the above mentioned ideology, We are fabricating an application using android design studio kit which is majorly concerned for the effortless access of all the essentials required in an institute.

Index Terms—Firebase Auth, xml, URL, API

I. INTRODUCTION

World is changing its pace at regular front, so as the technologies. Everyday, we are introduced with some or the other fresh development in the field of science and technology. Keeping a constant look over this changing world is a task in itself, but the services it bring us worth just like Mobile phones. Mobile phones have been evolved with stupendous operating systems throughout. One of the amazing operating system is Android. Android has been a fast developing operating system which has over taken other operating systems. The catchy feature of android is its versatility of several creations. It allows to develop simplest form of the application. The proposed application is one of the easiest approached system which is emphasised on the trouble-free piece of work engaged at institute on daily basis. Perfect use of digitalization trend can be witnessed through this application. Comfort has been the major concern while creation of this application. Accessing important documents, papers, video lectures and ofcourse newspapers just through a click is really very less tiring. Further sections will be focussed on the methodology and description of the technologies used while making this application.

II. LITERATURE REVIEW

In this [1] paper, class and staff Locator application has been developed via capstone. The Class And Staff Locator Application is used to locate classrooms and allocate staffs as per mentioned in their schedule. GPS (Global Positioning System) and Google Maps have been joined to this app, which plays the role of locating and tracking to reach desired destination. Location manager has been used for enabling the GPS system in java coding section which uses longitude and latitude of the device and hence enabling users to find current location of this application, map library has been used to convert longitude and latitude values into address of streets. Some Location based services intrudes Maps, MapActivity (a Location based API which is used to show location on the map). Map View has been used to display a view of map. MapActivity is used for controlling MapView. Location based API is used to locate user's current position and display that location on map and Google maps has been linked using Fragment Activity in coding section.

In this [2] paper, the development of a real-time response system has been done to intensify student's involvement in a class by means of keeping track of the attendance and participation of students. The lecturer keeps track of every student's attendance. This helps to enhance the participation of students and improves their attendance record. The access card feature has been developed and card is provided to every student and access is granted to get the teaching materials provided by the teachers. This is basically for the shy students those who are not very frank to the atmosphere. This way, teachers keeps a note of the progress of the whole class.

In this [3] paper, a test-based assessment system has been proposed. Students can take tests or can review their past tests on mobile devices, which reduces the hefty task of marking much number of test papers. Database of the proposed system consists of basic level tests for junior high school students in Taiwan. Teachers are having the benefit of composing digital test sheet by selecting questions from the given database. Teachers are also able to design their own questions. After the

completion of the test, test results can be obtained immediately by the teachers.

In this [4] paper, a mobile application has been developed for academic library which involves identifying services provided through library relevant to particular mobile user, development costs has been cut short by utilization of a number of open-source components, and results of the application has been evaluated. The features of the system includes: Library mobile website content, Mobile catalog, QR code scanner, and an Interactive library map.

In this [5] paper a feedback system for multiple user has been developed for a better learning in classrooms (rather cooperative). This system provides interactive mobile learning through a game-based approach to improve student's a collaborative learning. Students are learning to work as a team to accomplish a common aim, this is making them adaptive to the atmosphere and social development. The app has been intervened into different parts: Streaming camera, Media stream server, Screen recorder and Streaming video wall.

In this [6] paper, the Android application development challenge for colleges have been surveyed. This is an Android developer contest aimed at the college students in China. This contest had been held for two times since 2010. The main motive of this challenge is to motivate the college students to design and implement their applications using the Android platform. It gives students an opportunity to showcase their creativity and learn about the development of Android applications. The target of this contest is to keep the interest in the students intact rather increasing and to acknowledge other Students from other universities and regions about this contest so that an awareness about such development contest takes place and more number of students participate. The Android Application Development College Challenge is organised by Google to provide a platform to college students so that they can discuss and communicate the ideas and technology. The contest not only elevate student's creativity and practice, but also teaches about being a team player.

In this [7] paper, a mobile application named Fer Droid has been developed used for the collection of data from different students through this application. This application saves students precious time and gives an opportunity to contribute their time on other activities like studying, leisure, extra-curricular activities. This application is both subjective and objective as per the student's survey. The analysis of Fer Droid has concluded that students happy with this application and satisfied at the same time as they can retrieve information quicker than earlier by the help of this application.

In this [8] paper, a simple programming methodology has been developed which is based on MVP architecture to help the students to fabricate mobile applications in Android. The main idea of this is to design a set of methods, which is sorted and simple, so that any student can learn and develop a complete android app. The MVP architecture comprises of: model (that is enclosed with business objects), view (that comprises of all the UI components which makes the application and

forwards the interaction in between operations to the presenter) and presenter (that contains all the logic for the application).

In this [9] paper, an online physical simulation platform has been developed for Android programmers. Users those are basically students upload the Android applications to the server, which is connected to many mobiles. The screen of the phone is visible to the client because of the VNC Server. This paper has proposed an online platform for physical simulation in Android programming for the interested students. The online simulation platform consists of three layers: the Device Layer , it is a collection of embedded software running on smart phones, next is the Application Service Layer which is a bridge between the other two layers and it runs on an application server, next is the Operation Layer which is designed for the students living in remote areas, who can access the smart phone via a VNC Viewer client.

In this [10] paper, an university based application has been created which includes the university level stuffs those are list of holidays, syllabus, question papers and academic calender in the form of pdf that is stored into google drive for easy access. It includes another catchy feature which is "Ask your queries" which will let the users to clear any queries regarding the application usage or the material usage. Students are able to enjoy the easy access of the important requirements through this application. Application .apk file has been retrieved on the android device to access the application.

In this [11] paper, deals with the institute based requirements which help students to access the documents from a single place. Various forms and formats have been included in the application in the form of pdf and stored in google drive. Application size is also very nominal, won't acquire much space which is one of the advantages of this app. One exciting feature of this application is "feedback" portal, which lets the users to share their ratings and helps the developers to acknowledge the factual description of the application usage leading this app to improve and making it better.

III. METHODOLOGY

The system has been designed on android for keeping the versatility of the application building intact. This application is the summed up version of all the essentials for the graduation level students. System proposed is fabricated on android studio with the layout designing done on .xml extension. Students will be delighted to have all in one featured app. This application centralize itself on the effortless access of the question papers, video lectures, newspaper etc. at a single place. Students will be served with most comfortable access related to their education entitled with the university. As far as the technicalities are concerned, the features which catches eye are the video lectures linked in the app. itself. This is done through linking the API in the coding section. Parent-teacher portal has been designed and a separate forum has been fabricated for the interaction purpose. Newspapers have been added through their API source code (source code is must to avoid copyright issues).

A. Architecture

The above figure shows the flow chart of the proposed architecture.

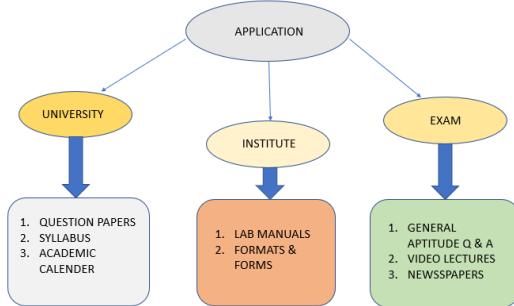


Fig. 1. System Architecture

Technologies used are described :

1.FIREBASE MESSAGING SERVICES: firebase messaging services has been used to synchronise application data across clients. It is a platform solution for messages and notification for android and without any cost. The coding has been shared :-

```

public class MyFirebaseMessagingService extends FirebaseMessagingService {
    private static final String TAG = "MyFirebaseMessagingService";
    @Override
    public void onMessageReceived(RemoteMessage remoteMessage) {
        Log.d(TAG, "From: " + remoteMessage.getFrom());
        if (remoteMessage.getData().size() > 0) {
            Log.d(TAG, "Message data payload is: " + remoteMessage.getData());
        }
        if (remoteMessage.getNotification() != null) {
            Log.d(TAG, "Notification Body: " + remoteMessage.getNotification().getBody());
        }
    }
}
  
```

Fig. 2. Coding firebase messaging service

2. NEWSPAPER LINKING: newspaper has been linked to the app through its API in coding section. Following are the results:

```

Intent intent = new Intent(Intent.ACTION_WEB_SEARCH);
intent.putExtra(SearchManager.QUERY, "http://timesofindia.indiatimes.com/");
startActivity(intent);
  
```

Fig. 3. Newspaper linking code

3. SEPARATE WEBSITE: application will contain separate website of its own for the parent-teacher interaction and other app related information will be included.

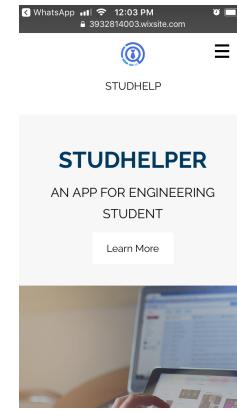


Fig. 4. Glimpse of Application website

IV. RESULTS

The experimental results are shown below:

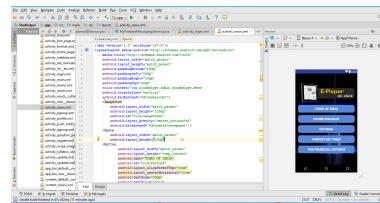


Fig. 5. News xml file



Fig. 6. Application screen view with institute section



Fig. 7. Snapshot of application of university section

V. CONCLUSION

This paper has been built in a way to make any reader understand the complete working of the application. This application will not only give students a platform of every useful documents but will keep them entertained by using the ask your queries block, video lectures and e-newspaper section. This application gives a complete exposure to the new and old questions, prior requirement of any student studying in any graduation school. Main moto of this application is user friendly and anyone can use it very easily moreover, this will capture only few MB of space in any android device.

REFERENCES

- [1] Ajay Shiv Sharma, Manpreet Singh Malhi, Manmeet Singh, Ramandeep Singh, CSLA – An application using Android Capstone Project ,*IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)*, 2014, pp 382-385.
- [2] Andi Sudjana Putra ,Ng Jun Jie, Tan Kok Kiong, Enhancing Student Involvement in a Class using Real-Time Response System*IEEE Transactions on Signal Processing*.
- [3] Wu-Ja Lin, Member, IEEE, Sin-Sin Jhuo, Chen-Hao Fan, Bo-Kai Ruan, and Po-Wen Chen, Test-based assessment with mobile devices ,*IEEE 17th International Symposium on Consumer Electronics (ISCE)*,2013,pp129-130.
- [4] Stan Kurkovsky, Wittawat Meesangnil,Building and Evaluating a Mobile Application for an Academic Library,*15th International Conference on Network-Based Information Systems*,2012,pp 357-363.
- [5] Chih-Tsan Chang, cheng-Yu Tsai, Song-En Peng, Pao-Ta Yu,hung-hsu tsai, On the Design of Multi-User Streaming Feedback System for Application of Cooperative Learning,*International Computer Symposium2016* ,pp 656-659.
- [6] Bin Peng, Jinming Yue, Chen Tianzhou, The Android Application Development College Challenge, *IEEE, 14th International Conference on High Performance Computing and Communications* 2012.
- [7] Hrvoje Maracic, Iva Bojic, Mario Kusek, Accessing Student Information Systems Using Mobile Connected Devices, *IEEE, EuroCon, Zagreb, Croatia* 1-4 July, 2013.
- [8] C. N. Ojeda-Guerra, A Simple Software Development Methodology Based on MVP for Android Applications in a Classroom Context, *IEEE, International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing 2015*.
- [9] Yu Liu, Ying Li, Jianwei Niu, Qinghua Cao, An online physical simulation platform for Android programming, *IEEE, Sixth International Conference on Technology for Education* 2014.
- [10] Aman Bansal, Rahul Pal, Md. Khaja Mohiddin, An Android Based Application Aimed at the University Related Forum, *stm journals, Recent Trends in Sensor Research and Technology* ISSN: 2393-8765, Volume 4, Issue 2, 2017.
- [11] Ishani Mukherjee, Mokshada Patra, Md. Khaja Mohiddin, An Interactive Application for the Facileness of Scholars, *stm journals, Journal of Telecommunication, Switching Systems and Networks* ISSN: 2454-6372, Volume 4, Issue 2, 2017.

VI. BIOGRAPHY



Ishani Mukherjee is pursuing Bachelor's degree in Electronics and Telecommunication Engineering

from Chhattisgarh Swami Vivekanand University, Bhilai from Bhilai Institute of Technology, Raipur, Chhattisgarh, India.



Mokshada Patra is pursuing Bachelor's degree in Electronics and Telecommunication Engineering from Chhattisgarh Swami Vivekanand University, Bhilai from Bhilai Institute of Technology, Raipur, Chhattisgarh, India.



Aman Bansal is pursuing Bachelor's degree in Electronics and Telecommunication Engineering from Chhattisgarh Swami Vivekanand University, Bhilai from Bhilai Institute of Technology, Raipur, Chhattisgarh, India.



Rahul Pal is pursuing Bachelor's degree in Electronics and Telecommunication Engineering from Chhattisgarh Swami Vivekanand University, Bhilai from Bhilai Institute of Technology, Raipur, Chhattisgarh, India.



Md. Khaja Mohiddin is pursuing his Ph.D. Degree in the field of Wireless Sensor Networks from GITAM University, Visakhapatnam, earned his M. Tech. Degree in the field of Digital Electronics and Communication Systems in 2012, earned his B. Tech. Degree in ECE in 2009 from Al-Ameer College of Engineering and Information Technology, Visakhapatnam affiliated to JNTU Kakinada. Then worked as a Lecturer for 1 year till 2010 in Al-Ameer College of Engineering and Information Technology, Visakhapatnam. Thereafter joined MATS School of Engineering, Raipur as Assistant Professor for 2 years till June 2012 and finally joined Bhilai Institute of Technology, Raipur as Senior Assistant Professor from July 2012 to till date. He is having 8.5 Years of Experience in the field of teaching. His Area of Expertise includes: Electromagnetic Waves, Antenna and Wave Propagation, Radar Engineering, Wireless Communication etc. Also having the technical knowledge related to Embedded Systems, Image Processing, Network Simulation etc.

Live Forensics Analysis Method For Random Access Memory On Laptop Devices

¹Danang Sri Yudhistira, ²Imam Riadi, ³Yudi Prayudi

¹Department of Informatics, Universitas Islam Indonesia

²Department of Information System, Universitas Ahmad Dahlan

³Department of Informatics, Universitas Islam Indonesia

Email: ¹danangsriyudhistira@gmail.com, ²imam.riadi@is.uad.ac.id, ³prayudi@uii.ac.id

Abstract— The development of computer technology now have an impact on the increasing cases of cybercrime crime that occurred either directly or indirectly. Cases of cybercrime now are able to steal digital information is sensitive and confidential. Such information may include email, user_id, and password. In addition to browser cookies stored on your computer or laptop hard drive, user_id, email, and password are also stored in random access memory (RAM). Random access memory (RAM) is volatile so that in doing the analysis required an appropriate and effective method. Digital data acquisition method in random access memory can be done live forensics or when the system is running. This is done because if the device or laptop computer is dead or shutdown then the information stored in random access memory will be lost. In this study has successfully carried out its acquisition of random access memory (RAM) for information access rights and password login form user_id on your websites such as Facebook, PayPal, internet banking, and bitcoin. Tools used to perform data acquisition, namely Linux Memory Extractor (LIME) and FTK Imager.

Keywords: Live forensics, RAM, laptop, devices.

I. INTRODUCTION

Along with the increasing use of computer crimes, then the chance of cybercrime is also increasing. Cybercrime is not only done to cripple a network server but also with steal critical data from an individual, organization or business entity [1]. Cases of cybercrime are happening now has already led to the theft of user_id or username, email, and password which is the nature of the personal information privacy for some people. Such information includes concerns about the facebook account, internet banking, PayPal, and bitcoin. User_id and password could be abused by people not responsible for how to steal other people's property and accounts result can harm the legitimate owner of the account [2]. In addition the result of theft and user_id password, could only occur if the label is a social media account stolen, whereas if the internet banking account is stolen then the possible occurred the crime of theft money transfer via internet banking to another account or by means of the imposition of a fee to the owner of the legitimate account if that account is used for illegal transactions with on behalf of the owner of legitimate internet banking account, however the address shipments addressed to the address of the thief [3].

Information in the form of email, user_id, and password in addition to browser cookies stored on is also stored in random

access memory on a laptop device that we use to do the login permissions. For it takes a proper method or technique, in order to perform the analysis of the random access memory on a laptop device. This is because the data in random access memory is volatile in nature. Volatile data will be lost if the computer is turned off or having to restart. Acquisition of random access memory to get information of digital evidence can only be done when the system is running or running [4].

Volatile data stored in random access memory describes the whole activity is taking place on a computer system that is being used. Handling data in random access memory has to be careful because in addition to its data can be lost if the system is turned off, the use of the tools will leave a footprint that can potentially overwrite existing valuable evidence is in the random access memory. It is, therefore, necessary the proper method for monetizing digital evidence stored in random access memory. The data acquisition method using the method live forensics [5].

Live forensics methods aimed at handling incidents faster, more assured data integrity, encrypted data can be opened and allow the memory capacity is lower when compared to traditional forensic methods. The stages are done in performing data analysis on random access memory with the live forensics method i.e., collect, examine, analyze and report [6][7].

II. LITERATURE REVIEW

In research conducted by (Anand, 2016) explained that the random access memory storing log-related information activities carried out by the user and the system is running [8].

Previous research conducted by (Nisbet, 2016) data acquisition device is done on a laptop is usually just for the acquired data information that exists on the hard disk but this time it could be done for the acquired data in random access memory. The focus of this study is to acquire data in random access memory to find files that are encrypted [9].

In research conducted by (Stüttgen, Vömel, & Denzel, 2015) explained that results from acquisitions in the random access memory we can see potential malware attacks. In addition, we can recognize malicious programs that are already installed on a computer operating system [10].

Research in methods of live forensics to analyze the random access memory requires accuracy in finding existing digital evidence. There is some hitch in the study of random access memory. For ease of handling the method live forensics as well as keeping the value of the integrity of the evidence, it will be accessible with the python scripting language (Bharath & R, 2015) [11].

Other research and development became the basis for the research is research conducted by (Divyang Rahevar, 2013). On the research tells us that hidden file-related information, user_id, password, rootkits, and sockets are not only stored on the hard disk but also stored in random access memory [12].

In research conducted by (Richard Carbone, 2012), conducted 2 tools namely LiME and Fmem comparison to get results in the acquisition random access memory on the pc-based Linux operating system by using the framework volatility memory analysis [13].

Other research conducted by (Karayianni & Katos, 2012) and investigations about data privacy regarding the information which is personal data. The information in the form of passwords stored in random access memory. Process analysis of the random access memory when the computer is done in conditions of operation or running because if the computers in a dead condition then data stored in random access memory will disappear [14].

III. CURRENT PRACTICES

A. Linux Memory Extractor (LiME)

Linux Memory Extractor (LiME) tapes loadable kernel module that is can be used to perform the acquisition of volatile memory on Linux based-powered device. To be able to use the LiME needed privilege as root. LiME is the first tools capable of capture random access memory as a whole [15].

B. FTK Imager

FTK Imager or complete language is "Forensic Toolkit Imager" is a stand-alone application for the hosts Access disk imaging Data. Access to Data is a company engaged in the field of digital forensics and provides solutions from a classroom stand-alone to enterprise-class for digital investigation process. FTK Imager tools are used to analyze the results of the data acquisition of the laptop-based Linux operating system [16].

IV. ACQUISITION OF RANDOM ACCESS MEMORY

The source of data used for digital evidence in this study comes from the random access memory on the laptop-based Linux operating system. In making acquisitions in the random access memory there are several stages that are done as in figure 1.

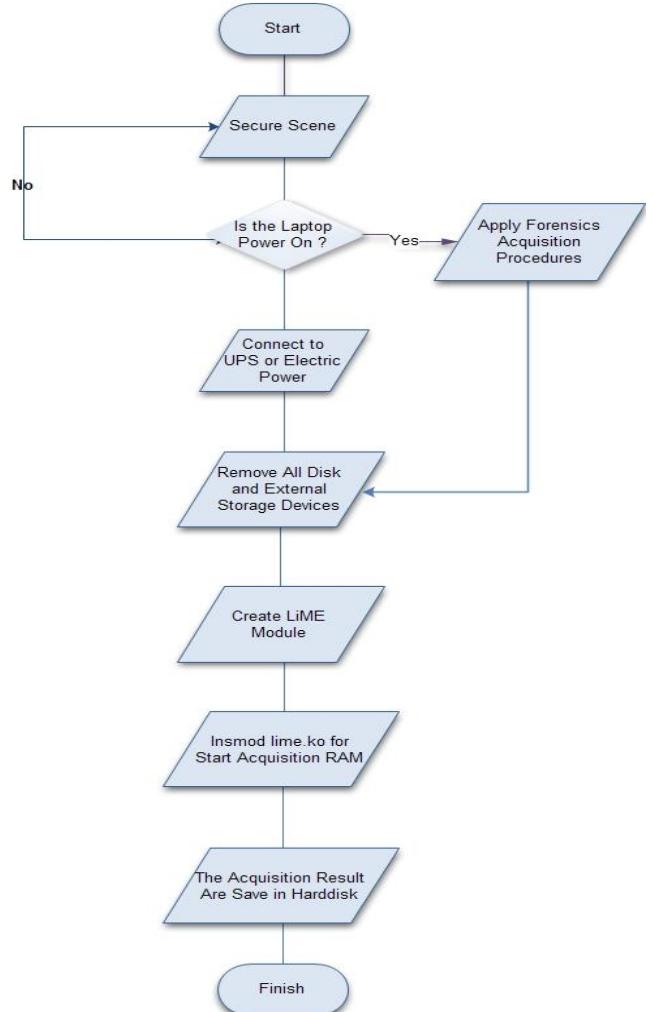


Figure 1. Flowchart Aquisition Of Random Access Memory

In Figure 1 that the first process begins with securing the scene of things. If the laptop is turned on then it can be done the next process, namely the acquisition of random access memory. Do not ever take off the power cable so that the laptop did not die when it conducted the process data acquisition. Prior to the acquisition process begins, remove all external storage that is stuck to the laptop. Create module LiME (Linux Memory Extractor) and type the command instead to begin the process of capture random access memory. Proceeds from the sale to random access memory will be stored automatically in the folder directory/home.

To create the LiME module (Linux Memory Extractor) first we go to the directory/src on the LiME-master folder located in folder Download by typing commands in accordance with Figure 2.

```

root@lepiku:/home/santoku/Downloads# cd LiME-master
root@lepiku:/home/santoku/Downloads/LiME-master# ls
doc LICENSE README.md src
root@lepiku:/home/santoku/Downloads/LiME-master# cd src
root@lepiku:/home/santoku/Downloads/LiME-master/src# ls
disk.c           lime.h   lime.o Makefile      Module.symvers
disk.o           lime.mod.c main.c Makefile.sample  tcp.c
lime-4.4.0-31-generic.ko lime.mod.o main.o modules.order  tcp.o

```

Figure 2. Access of LiME On The Directory SRC

Figure 2 describes the process to access the Linux Memory module Extractor (LiME) on the Linux operating system-based laptop. The process is done in a live forensics or laptop in the condition turned on.

The next process to capture the memory on a random access memory by typing commands at the command line terminal Linux fits in Figure 3

```
root@lepiku:/home/santoku/Downloads/LiME-master/src# insmod lime-4.4.0-31-generic.ko "path=/home/santoku/skenario-linux.lime format=lime"
root@lepiku:/home/santoku/Downloads/LiME-master/src#
```

Figure 3. Order to Capture Of Random Access Memory

Figure 3 describes the process of capture random access memory on Linux laptop device. Capture process starts with access to the directory/SRC to find the kernel module lime-4.4.0-31-generic.ko. Next, do the process of capture by typing commands at the command line Linux terminal according to Figure 3. Proceeds from the sale to random access memory on the laptop-based Linux operating system will be stored in the directory/home.

Proceeds from the sale to random access memory there is a file with the extension *.lime. This file can be analyzed using tools FTK Imager that runs the Windows operating system. Analysis of the results obtained some information that can be used as evidence. Such information concerns the user_id and password that originates from the internet banking account, PayPal, Bitcoin, and Facebook.

V. THE ANALYSIS OF RANDOM ACCESS MEMORY

Based on the results of the acquisition of the random access memory on the laptop-based Linux operating system there is some related information that can be used as evidence. The information of which is information that is confidential or privacy because it is a personal account belonging to someone who used to do the login permissions an application on the website.

In this research have successfully found evidence in the form of digital information user_id and password used to access the internet banking login, PayPal, Bitcoin and facebook. The information clearly captured by tools Linux Memory Extractor (LiME) and able to be analyzed using tools FTK Imager.

The first successful evidence obtained is the user_id and password to access login to facebook website page. Proof of this is in addition to stored in browser cookies are also stored in random access memory devices. Evidence of the user_id and password facebook listed in Figure 4.

14 01 AD 14 01 01 68 74-74 70 73 3A 2F 2F 77 77 ..-. https://www
77 2E 66 61 63 65 62 6F-6F 6B 2E 63 6F 6D 2F 6C w.facebook.com/1
6F 67 69 6E 2E 70 68 70-68 74 74 70 73 3A 2F 2F egin.phphttps://
77 77 77 2E 66 61 63 65-62 6F 6F 6B 2E 63 6F 6D www.facebook.com/
2F 6C 6F 67 69 6E 2E 70-68 70 65 6D 61 69 6C 6D /login.phpmailto
61 6E 7A 79 75 64 68 69-73 74 69 72 61 40 72 66 anzyudhistira@ro
63 6B 65 74 6D 61 69 6C-2E 63 6F 6D 70 61 73 ckemail.compass
6A 6F 67 6A 61 31 32 33-35 37 31 68 74 74 70 73 jogja12351https
3A 2F 2F 77 77 77 2E 66-61 63 65 62 6F 6F 6B 2F //www.facebook.
63 6F 6D 2F 01 01 5A 64-C4 DF 00 00 00 00 00 00 com.zdbs...
00 00 40 OB 00 00 01 00-00 00 0A 00 00 00 6C 00 ..@.....
6F 00 67 00 69 00 6E 00-5F 00 66 00 6F 00 72 00 o-g-i_n_-f-o-r
6D 00 04 00 00 70 00-6F 00 73 00 74 00 C1 06 m...-p-o-s-t-A
00 00 68 74 74 70 73 3A-2F 2F 77 77 77 2E 66 61 ..-https://www.fa
63 65 62 6F 6F 6B 63-6F 6D 2F 6C 6F 67 69 6E cebook.com/login
2E 70 68 70 3F 73 6B 69-70 5F 61 70 69 5F 6C 6F .phoskip.apilic

Figure 4. Evidence Of The Facebook Account

Based on the analysis of the results of the acquisition of random access memory that is listed in Figure 4, note that proof of access logs in facebook using the email account "danzyudhistira@rocketmail.com" and the password "jogja123571". This evidence will still be stored in random access memory and will not be lost as long as the laptop is not turned off.

Other evidence that successfully obtained i.e. access login to your website belongs to national banking. Access to internet banking transaction. Evidence from access the login listed in Figure 5.

Figure 5. Evidence Of The Internet Banking Account

Based on the analysis of the results of the acquisition of random access memory that is listed in Figure 5, it is noted that proof of access to the internet banking account login using user _ id "danangsr1911" and the password "123571". The account is stored in random access memory without experiencing encryption. This proves that the internet banking application belongs to the national banking still vulnerable to acts of theft user_id and password by the person who is not liable if it managed to get the account access login.

In addition, there is evidence of the access account login bitcoin. Bitcoin is a cryptocurrency currency now is quite popular to use for online transactions. Bitcoin account login access evidence listed in Figure 6.

Figure 6. Evidence Of The Bitcoin Account

Based on the analysis of the results of the acquisition of the random access memory on a laptop device as listed in Figure 6, note that the proof of the access account login using bitcoin email "danangsriyudhistira@gmail.com" and the password "Yogyakarta123571". Just as the internet banking account, the account is also not experiencing bitcoin encryption.

Other evidence of the successful results obtained from sale to random access memory on the device operating system Linux-based laptop that is PayPal account login access. A PayPal account is also often used for payment transactions online. PayPal applies internationally making it easy to use just about any transaction. Paypal account login access evidence listed in Figure 7.

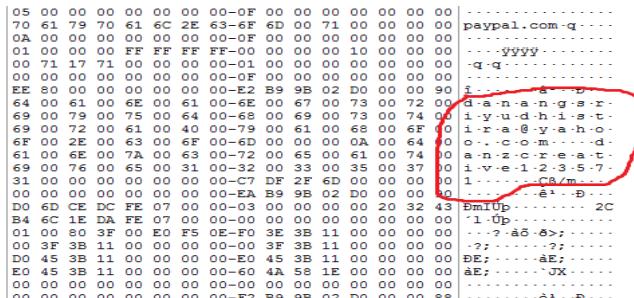


Figure 7. Evidence Of The Paypal Account

Based on the analysis of the results of the acquisition of the random access memory on a laptop device as listed in Figure 7, note that evidence access login your PayPal account using the email "danangsriyudhistira@yahoo.com" and the password "danzcreative123571". This evidence will still be stored in random access memory for laptop device is not turned off. Just as the internet banking accounts and user_id and password bitcoin, PayPal account stored in random access memory is also not experiencing the encryption

VI. CONCLUSION

After a series of research and analysis of random access memory on the device, a laptop with Linux operating system using the method live forensics can be drawn the conclusion that random access memory capable of storing all the information all the related activities performed by the user or users. In this case, it is activities to access internet banking login, PayPal, Bitcoin, and Facebook. Proof of access the login i.e. user_id and password. Tools Linux Memory Extractor (LiME) able to do capture memory thoroughly so that the information obtained from random access memory was able to complete and can be used as evidence in a digital handling of crimes and involve evidence-based laptop Linux operating systems. While the FTK Imager tools are able to perform analysis of digital evidence properly because of evidence that encrypted not encrypted are also capable opened by these tools.

VII. FUTURE WORK

This research has managed to find a social media account login access facebook, internet banking, bitcoin and PayPal stored in random access memory on the device a laptop either user_id or username and password. For the development of further research is expected to find a credit card account that was inputted when we conduct E-Commerce transactions to shop online using a browser laptop. In addition to the credit card account is saved in the browser's cookies will also be

stored in random access memory on a device that is used for the transaction, in this case, using a laptop devices

REFERENCES

- [1] Wahyudi, E., Indonesia, U. I., Riadi, I., Dahlan, U. A., Pray, Y., & Indonesia, U. I. (2018). Virtual Machine Forensic Analysis And Recovery Method For Recovery And Analysis Digital Evidence. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(2), 1–7.
- [2] Prayogo, A., Riadi, I., & Luthfi, A. (2017). Mobile Forensics Development of Mobile Banking Application using Static Forensic. *International Journal of Computer Applications*, 160(1), 5–10.
- [3] Riadi, I., & Umar, R. (2017). Identification Of Digital Evidence On Android ' s. *International Journal of Computer Science and Information Security*, 15(5), 3–8.
- [4] Dave, R., Mistry, N. R., & Dahiya, M. S. (2014). Volatile Memory Based Forensic Artifacts & Analysis. *International Journal For Research In Applied Science and Engineering Technology*, 2(I), 120–124.
- [5] Rochmadi, T., Riadi, I., & Prayudi, Y. (2017). Live Forensics for Anti-Forensics Analysis on Private Portable Web Browser. *International Journal of Computer Applications (IJCA)*, 164(8), 31–37.
- [6] Riadi, I., Eko, J., Ashari, A., & -, S. (2013). Internet Forensics Framework Based-on Clustering. *International Journal of Advanced Computer Science and Applications*, 4(12), 115–123.
- [7] Umar, R., Riadi, I., & Zamroni, G. maulana. (2017). A Comparative Study of Forensic Tools for WhatsApp Analysis using NIST Measurements. *International Journal of Advanced Computer Science and Applications*, 8(12), 69–75.
- [8] Anand, V. N. (2016). Acquisition Of Volatile Data From Linux System. *International Journal of Advanced Research Trends in Engineering and Technology*, 3(5), 95–97.
- [9] Nisbet, A. (2016). Memory forensic data recovery utilising RAM cooling methods, 11–16.
- [10] Stütten, J., Vömel, S., & Denzel, M. (2015). Acquisition and analysis of compromised firmware using memory forensics. *DFRWS 2015 Europe*, 12(S1), S50–S60.
- [11] Bharath, B., & R, N. M. A. (2015). Automated Live Forensics Analysis for Volatile Data Acquisition. *Int. Journal of Engineering Research and Applications*, 5(3), 81–84.
- [12] Divyang Rahevar. (2013). Study on Live analysis of Windows Physical Memory. *IOSR Journal of Computer*

Engineering (IOSR-JCE) , 15(4), 76–80.

- [13] Richard Carbone. (2012). The definitive guide to Linux-based live memory acquisition tools. *DRDC Valcartier TM 2012-319*.
- [14] Karayianni, S., & Katos, V. (2012). Practical password harvesting from volatile memory. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 99 LNICST*(May 2014), 17–22.
- [15] LiME, <https://github.com/504ensicsLabs/LiME>, 2018
- [16] FTK Imager, <https://accessdata.com>, 2018

Evaluation of Snort using rules for DARPA 1999 dataset

Ayushi Chahal¹, Dr. Ritu nagpal²

Department of Computer Science and Engineering,
Guru Jambheshwar University of Science & Technology,
Hisar, India.

¹ayushichahal@gmail.com,²ritu_nagpal22@yahoo.co.in

Abstract— Network security is main concern now-a-days and Snort is one of the advanced techniques that is used to tackle rising security threats over the internet. Snort is kind of Network Intrusion Detection System that allows user to write their own rules to detect different attacks over the network on the basis of their signatures and further gives freedom to users to handle these attacks in different ways. MIT-DARPA 1999 dataset (which consists of both normal and abnormal traffic) is used for evaluation of Snort in this paper. This evaluation is done with the help of the proposed detection rules. In this paper we have detected all kinds of attacks i.e. DOS, U2R, R2L, Probe and data attacks. These proposed rules results are further compared to the Detection Scoring Truth of DARPA 1999 dataset.

Index Terms— Network Intrusion Detection System (NIDS); Snort; detection rules; DARPA dataset.

I. INTRODUCTION

Researches about network intrusion and intrusion detection began in the early 1980s by James Anderson. Intrusion detection system (IDS) in the recent years is generally considered to be the second line of defense after the firewall. IDS provide real-time protection ability and can intercept the invasion before the whole network system is endangered [1]. Different kinds of IDS are present in the market for detecting and protecting the data packet traffic over the network, like: Network based IDS (NIDS), Host based IDS (HIDS).

A. SNORT-IDS

Snort is one of the famous and most effectively used NIDS against intruders.[2] It is signature based NIDS. Snort is an open source network intrusion prevention and detection utilizing a rule-driven language. Snort is available under the GNU (General Public License) [3].

Snort is a cross-platform operating system developed by Martin Roesch in 1998 [4]. He further found Sourcefire in

2001 [5] which created a commercial version of Snort having GPLv2+. With the acquisition of Sourcefire in October 2013, Snort is now one of the technologies used in Cisco products.

Snort++ is an updated version of the Snort IPS (intrusion prevention system).It uses friendly design with build in documentation and configuration. In it command shell allows interaction with running instance of snort. It provides facility of Auto-Detection of all protocols on all ports. It support multiple packet processing threads. It uses simplified rule language. It support sticky buffers in rules. It auto-detect services for portless configuration. It makes key components pluggable. [15]

B. DARPA dataset

DARPA dataset is of interest to all researchers working on intrusion detection. It had been created by MIT Lincoln Laboratory IDS evaluation methodology. Such evolution was carried out in 1998 and 1999 which result out in form of: “1998 DARPA Intrusion Detection Evaluation Data Sets” and “1999 DARPA Intrusion Detection Evaluation Data Sets”, “2000 DARPA Intrusion Detection Scenario-Specific Datasets” of experiments run in 2000.

1) DARPA dataset 1999

First DARPA dataset 1998 came into existence, after some evolution in it 1999 DARPA Intrusion Detection became the dataset of interest to all the researchers. For evaluation of dataset DARPA 1999, it is divided into two parts [6]:

- Real-time Evaluation
- Off-line Evaluation

IDSs were tested as a part of real-time evaluation, off-line evaluation or both.Data collected under DARPA 1999 is of five weeks. Over all data inside this dataset is considered under two phases i.e. **Training dataset** and **Testing dataset**.

First three week data are Training data. In 1999, IDSs were trained with the help of dataset 1998 as well as dataset of 1999. Fourth week dataset and fifth week dataset are used as Testing data.

DARPA 1999 dataset contains very limited number of attacks that are detectable with the fixed signature. These attacks are divided into four categories. Detail of every category of attack is as follows:

1. **DOS (Denial Of Service) attack**: This type of attack occurs when legitimate users are not able to use computing and memory resources because intruder makes these resources too busy to handle authorized requests. Different kind of DOS attacks are shown below by table 1:

Table 1: Different kind of DOS attacks

I. Apache	II. selfping
III. appoison	IV. Smurf
V. Back	VI. sshprocesstable
VII. crashiis	VIII. syslogd
IX. dosnuke	X. tcprerset
XI. Land	XII. Teardrop
XIII. mailbomb	XIV. udpstrom
XV. SYN Flood	XVI. warezmaster
KVII. Ping Of Death	VIII. warezclient
XIX. Process Table	

2. **Remote to Local (R2L) attack**: In it attacker who do not have account on the remote node sends packet to that node through the network. Attacker then exploits some vulnerability to gain access as a local user to that remote node. There are very difficult to detect as they involve both network level features such as “duration of connection” and “service requested” and host level features like “number of failed login attempts”. Different kind of R2L attacks are given below by table 2:

Table 2 : Different kind of R2L attacks

I. Dictionary	II. ftp-write
III. Guest	IV. htptunnel
V. imap	VI. Named

VII. ncftp	VIII. netbus
IX. netcat	X. phf
XI. ppmacro	XII. sendmail
XIII. sshtrojan	XIV. xlock

3. **User to Local (U2R) attack** : The attacker starts as a normal user on the system and becomes root user by gaining the root access through exploiting vulnerabilities. It involves exploitation on semantic details that's why these attacks are difficult to capture at early stage. It features such as “number of file creations” and “number of shell prompts invoked”. Different kind of U2R attacks are shown by table 3 below:

Table 3 : Different kind of U2R attack

I. anypw	II. casesen
III. Eject	IV. ffbconfig
V. fdformat	VI. loadmodule
VII. ntfsdos	VIII. Perl
IX. Ps	X. sechole
XI. xterm	XII. yaga

4. **Probe attack**: The attacker scans the network of computers to collect information or to find known vulnerabilities. The attacker can use this information to exploit the nodes over this network. Hence, basic connection level features such as the “duration of connection” and “source bytes” are significant while features like “number of files creations” and number of files accessed” are not expected to provide information for detecting probes. Different kind of Probe attacks are shown by table 4 below:

Table 4 : Different kind of Probe attack

I. insidesniffer	II. ipsweep
III. is_domain	IV. mscan
V. ntinfoscan	VI. nmap
VII. quesco	VIII. resetscan
IX. Saint	X. Satan

2) DARPA dataset 2000

After improvement in the DARPA 1999 dataset, DARPA 2000 has evolved. DARPA 2000 is a simulated network which is divided into three segments:

- 1 network outside Air Force Base (AFB)
- 2 network inside AFB
- 3 DMZ network which connects both inside and outside networks of AFB.

It includes two attack scenarios: LLDOS1.0 and LLDOS2.0.2. [6]

II. BACKGROUND AND RELATED WORK

Martin Roesch gave the clear-cut difference between snort and tpdump. He stated the basic working model of snort. According to him, snort consists of 3 primary subsystem: *the packet decoder, the detection engine and the logging & alerting system*. He gave a simple way to write snort rules. It has become a small, flexible and highly capable system that is used all around the world on both small and large scale network. He also define some applications of snort in his paper [4] like Snort can be used to characterize the signature of the attack, Snort can be used as Honey-pot monitors, Snort can help in “focused monitoring”.

Extended form of Snort architecture is presented by Kurundkar G.D *et al.*[7] as they gave snort component architecture with six components, namely : *Packet Decoder, Preprocessor, The Detection Engine, Logging and Alerting System, Output Modules*. They have explained different type of intrusion detection system like NIDS (Network Intrusion Detection and Prevention System), NBA(Network Behavior Analysis System), HIDS(Host Intrusion Detection System) and IDPS(Intrusion Detection and Prevention System). IDPS provide multiple detection methods: Signature based, Statistical Anomaly based, Stateful Protocol Analysis IDPS.

The best approach to any organization to perform penetration testing is to write snort rules to protect against attack. There is no alternative of secure coding. Alaa El- Din Riad *et al.*[8] presented a new frame work that is designed with using data visualization technique by using Jquery & php for analysis and visualizes snort result data for user. They presented a new way to represent Snort rule in form of rule header and rule option. Rule header contains information about what step should be taken if all the content in rule option matches.

Intrusion Detection rules such as Snort rules are increasingly becoming complicated and massive these days.[1] presented an innovative way which will largely

enhance the detection efficiency in both space and time aspect. It gave an innovative way to organize rule in form of a three dimensional list. In it rule is firstly divided into various categories like alert, logs, or pass type and then further compartmentalized by different protocol types which is further divided into RTN (Rule Tree Node) i.e. header and their two pointers and OTN (Option Tree Node) i.e. body section and pointer to other body section, as that was divided in Two-dimensional implementation.

A new method for driving and testing malicious behavior of detection rule is introduced by Raimo Hilden *et.al.* which group the rules by their shared contents and extracts the most general rule of each group to optimized rule base. These pruned rules are stored for diagnostics. This method also maintains logs of which rule belong to which generalization. [9] uses rule generalization and rule base optimization. Key relation between signatures, rules and traffic packets must be maintained to avoid false positives and negatives in rule engine. It used content descriptor, connection descriptor, Match function to optimize the any rule R. It used signature parser, signature verifier, Signature Syntax Warning Logs, Rule Generator, Rule Verifier, Rule Syntax Warning Logs, Rule Purifier, Substitution Table in its architecture. For testing its results it used two methods: *dummy method, novel method*. In dummy method, pruning was done only on identical rules, and there was no proper linkage between rules and signature. In novel method, each file is given a signature with a unique identifier, in it they verified syntax of each signature. It greatly assists the experts in work as now they do not need to manually inspect signature, rules and traffic packet, they can now concentrate on automatically detected and reported issues.

Snort works on the signatures usually engineered based on experience and expert knowledge. It requires long development time. [10] gave approaches for an automated re-use of design of existing signatures. Sebastian Schmerl *et al.* showed their re-use approach for single-step signatures used by Snort. After selecting the related signatures with the help of signature generalization called abstraction, the engineer can look for similarities. Abstraction can be accomplished by iterative applying Transformations. This will result in Abstraction Tree. Each Transformation is weighted by a metric which defines similarity. The signature with lowest abstraction degrees are selected to understand nature of attack.

To improve Snort rules for Probe attacks N. Khamphakdee *et al.* uses MIT-DARPA 1999 data set. Firstly analysis of existing Snort-IDS rules was done to improve the proposed Snort-IDS rules. Secondly, Wireshark software was applied to analyze data packets form of attack in data set. Finally, the Snort-IDS was improved, and it can detect the

network probe attack. In [11] Probe attacks are divided into six types based on its nature. Comparison of Snort rules with the Detection Scoring Truth on the basis of efficacy of detection attacks is done and results of the tested Snort-IDS rules confirm that the proposed Snort-IDS can correctly detect 100% of the network probe attacks. Regarding to the comparative analysis with the notification Detection Scoring Truth, the detection number of the proposed Snort-IDS rules are more than the detection scoring truth. Because, some moments of the attacks had occurred in several times and attack occur in several time but the Detection Scoring Truth identify as one time.

[12] has shown that any rule that has one or more content matches in it has a fast pattern associated with it .The string that Snort puts into its fast pattern matching engine to begin the process of detection is chosen somewhat intelligently by Snort itself. This pattern is usually the longest string in a rule, because the longer the string is, the faster a rule will be. The goal of a rule-writer should be to choose a fast pattern, if one can generate an alert for most of the times when he/she enter a rule, then he/she has successfully targeted his/her detection, and written a rule with the minimum possible performance impact on Snort.

Most of the IDS researchers prefer to work on DARPA dataset with Snort. S. Terry *et al.* used tcpdump files as input to the Snort which was configured with all rules and alert file was produced for each tcpdump file. He analyzed performance of Snort with DARPA dataset 1999 on the basis of number of malicious connection detected. [13] Shows that for *DOS attack*, Snort performed best on back and land attack while it does not perform up to the mark for Smurf, Syslog and teardrop attack. In case of *R2L attack*, Snort performed best in case of phf attack with no false positive and performs worst in case of spy, warez, ftp, warezclient, warezmaster attack with low true positive rate. Snort did not performed that well for *U2R attack* as compared to other three attacks, yet it gives impressive results by reducing few false positives. In case of *probe attack*, threshold for detecting these attacks is very low in corresponding Snort rules and hence, majority of ipsweep connections was detected.

A. Saboor *et al.* evaluate Snort against DARPA attack DDoS attack with different hardware configurations. He used three test benches with different configurations but having Linux installed on each test bench as operation system.

- In these test benches, he used Hping for DDoS attack simulation tool, Hping and Ostinato for background traffic generation tool, Hping and Teamviwer for DoS verification tools.

- For attack, he used two scenarios i.e. attack scenario 1 having attack traffic only and attack scenario 2 having mix traffic.
- For evaluation matrices, he used maximum packet rate, resource availability, throughput, Snort rate filter option.

From his experiment [14] he concluded that in terms of detection capability of Snort no test bench outperformed other due to lack of Snort signature database and rate filtration. Using more RAM improved the performance of Snort. Hardware implementation decreased drop of packet to atmost 50%, but it did not show any improvement in packet handling and attack detection capability of Snort.

III. CREATING SNORT RULES

A. Experimental Setup

Rules are designed to detect different type of attacks with the help of their signatures. These rules are made with the help of Snort 2.9.8.3 with DAQ 2.0.6 version. It operated on Ubuntu 14.04 LTS operating system using Intel Pentium processor.

B. SNORT installations

We install snort in Ubuntu in root directory with the help of installation manual of sublime roots. At the same time, we also install DAQ library in root.

First we install pre-requisites for snort like libpcap, libdumbnet, libpcrc etc. Then we install and configure snort. After that, we create some directories like /etc/snort to configure and write rules, /var/log/snort which is used to store alerts generated during process. Also we adjusted paths of the local rules and white lists and black list rules.

C. Dataset

DARPA 1999 dataset is used to make rules. As described above, DARPA 1999 have five weeks data. From which week 1, week 2, week 3 are training dataset. Since, we work on NIDS so we opted for week 2 dataset as a training dataset, because only this dataset contains labeled attacks. So with the help of week 2 dataset we made different rules. These rules are then tested on the testing dataset i.e. week 4 and week 5 dataset, to get our results. We have utilized .tcpdump file format by choosing *inside.tcpdump* and *outside.tcpdump* files in week 4 and week 5.

D. Proposed SNORT rules

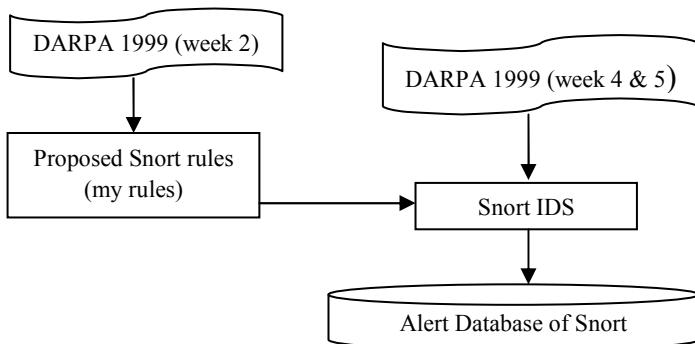


Fig 1 : Flow diagram shows making of SNORT rules

Week 2 dataset consists of some labeled attacks with their signatures inside, while week 1 and week 3 do not have any labeled attacks. So dataset like week 1 and week 3 can be used for IDS like anomaly detection. Since, snort is a NIDS and can detect attack on the basis of some content or signature of the attack inside the packet. With the help of these signatures of different attacks, different rules are made.

Hence we have used week 2 dataset as training data to create our rules. We took some signatures for each attack type and each attack type is given its corresponding *classtype* in which it fits best. Every rule is assigned a unique *sid* number.

We have generalized these rules by using “*any*” keyword in place of source and destination address places as well as for source and destination Port address places, so that they can send alert for all kind of Source and Destination addresses and port numbers. Some of the rules which are used to get final results are shown below:-

```

alert udp any any > any any (msg:"NTinfoscan attack which is kind of Probe
is detected"; content:"NTinfoscan"; nocase; sid:10000001; rev:01;
reference:url, https://www.ll.mit.edu; classtype:attempted
recon;)

alert udp any any > any any (msg:"pod attack which is a type of DOS is
detected"; content:"pod"; nocase; sid:10000002; rev:01; reference:url,
https://www.ll.mit.edu; classtype:denial-of-service;)

alert udp any any > any any (msg:"back attack which is type of DOS attack is
detected"; content:"back"; nocase; sid:10000003; rev:01;
reference:url, https://www.ll.mit.edu; classtype:denial-of-service;)

alert icmp any any > any any (msg:"ipsweep attack which is kind of Probe
attack is detected"; content:"00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00";
nocase; sid:10000049; rev:01; reference:url, https://www.ll.mit.edu; classtype:icmp
event;)
  
```

```

alert tcp any any > any any (msg:"NTinfoscan attack which is kind of Probe
attack is detected"; content:"NTinfoscan"; nocase; sid:10000017; rev:01;
reference:url, https://www.ll.mit.edu; classtype:tcp-connection;)

alert tcp any any > any any (msg:"pod attack which is a type of DOS attack is
detected"; content:"pod"; nocase; sid:10000018; rev:01; reference:url,
https://www.ll.mit.edu; classtype:denial-of-service;)

alert tcp any any > any any (msg:"back attack which is type of DOS is attack
detected"; content:"back"; nocase; sid:10000019; rev:01; reference:url,
https://www.ll.mit.edu; classtype:denial-of-service;)

alert tcp any any > any any (msg:"crashiiis attack which is kind of DOS attack
is detected"; content:"crashiiis"; nocase; sid:10000051; rev:01; reference:url,
https://www.ll.mit.edu; classtype:denial-of-service;)

alert ip any any > any any (msg:"NTinfoscan attack which is kind of Probe
attack is detected"; content:"NTinfoscan"; nocase; sid:10000033; rev:01;
reference:url, https://www.ll.mit.edu; classtype:tcp-connection;)

alert ip any any > any any (msg:"pod attack which is a type of DOS attack is
detected"; content:"pod"; nocase; sid:10000034; rev:01; reference:url,
https://www.ll.mit.edu; classtype:denial-of-service;)
  
```

IV. PERFORMANCE EVALUATION

This section describes the experimental evaluation of the SNORT rules which are created to compare the detection performance.

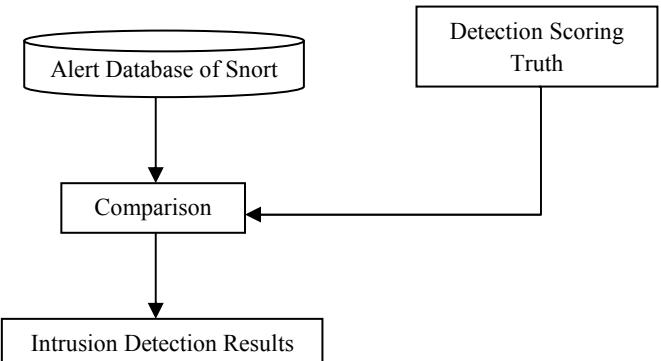


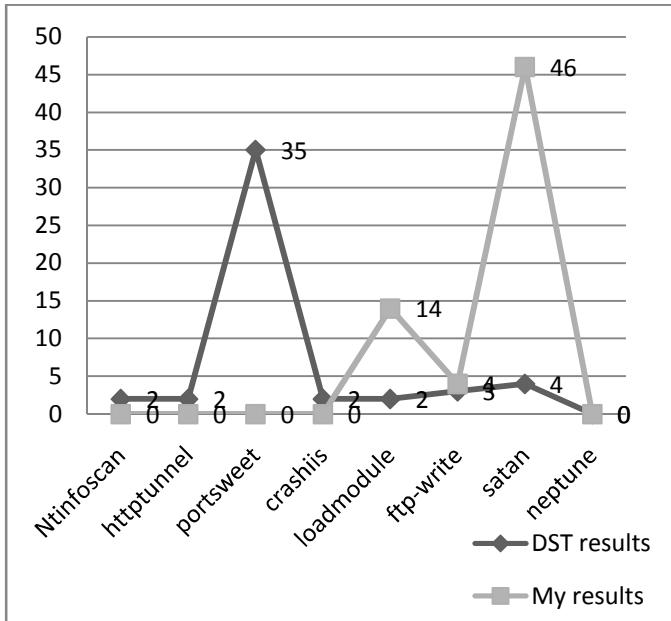
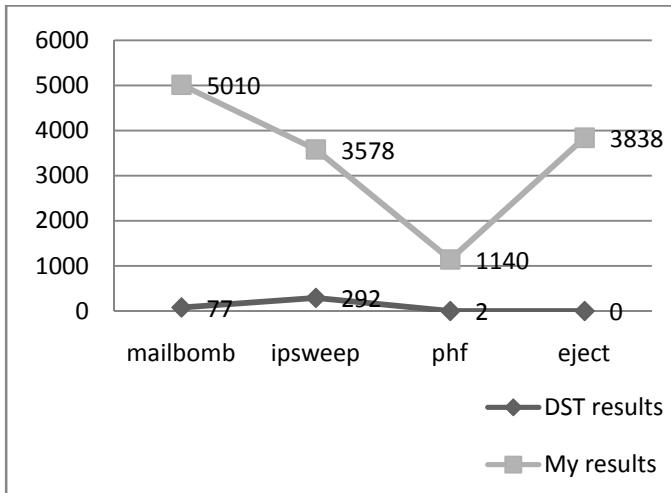
Fig 2: results of newly created rules compared with detection Scoring Truth

Figure 2 is a flow diagram shows the procedure of attack detection comparison of created rules with Detection Scoring Truth. Notifications that we get in alert database are compared with actual attack in Detection Scoring Truth.

We have used different type of attacks like Dos, U2R, R2L, Probe attacks to get our results and evaluate the performance. Here are the results shown in the form of graph. Graph is between the number of alerts we get through the SNORT rules (when applied on Week 4 and Week 5 data set of DARPA 1999) and Detection Scoring Truth.

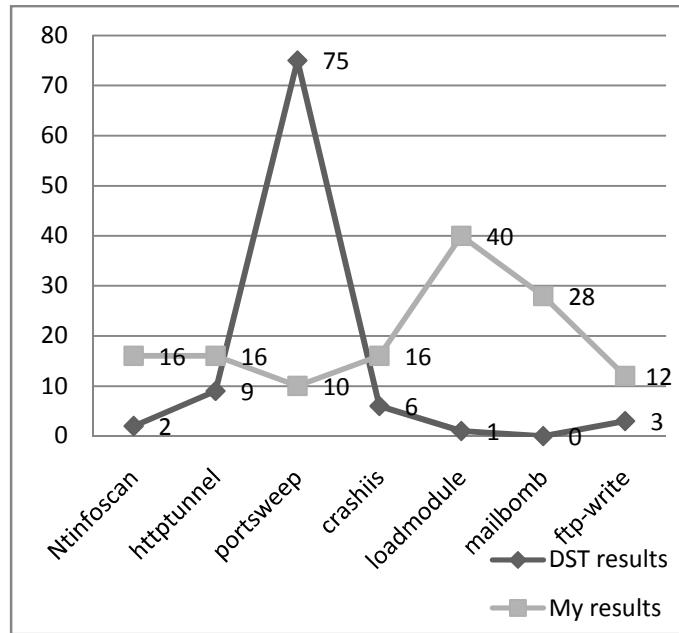
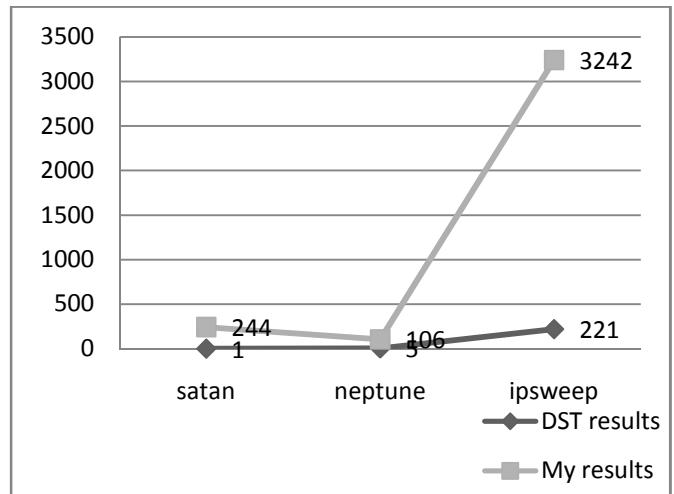
1. Comparative results of week 4 with Detection Scoring Truth (DST) :

Here, in the graphs shown below result we can see that Data Scoring Truth of DARPA 1999 detects more attacks for NTinfoscan, httptunnel, portsweep, crashiis attack but for loadmodule, ftp-write, satan my results are far better than that of DST results.



2. Comparative results of week 5 with Detection Scoring Truth (DST) :

Except portsweep attack all the remaining attack's rules in above graph namely NTinfoscan, httptunnel, crashiis, loadmodule, mailbomb, ftp-write performed well and give good results by giving appropriate amount alerts after detecting corresponding attacks.



V. CONCLUSION AND FUTURE DIRECTIONS

We have created some Snort rules that are used to detect these signature based attacks. These rules also classify attacks according to their characteristics into different classtypes. DARPA dataset is considered as dataset of interest for intrusion detection researchers. So, we used DARPA training dataset to create Snort rules for different attacks. These rules are used in generalized form so that it can maximize the alert detection. This generalization is necessary so that rules can detect novel attacks. This generalization is achieved by relaxing some of the conditions on the rule. These rules are then tested on DARPA testing dataset which provide good results when compared to Detection Scoring Truth of DARPA

dataset. Overall proposed rules performed very well as compared to DST rules.

There are several future directions for research:

First, there is need to improvise these rule with respect to false alerts. As we have discussed different factors for reducing false alerts like rule generalization, clustering method, alert correlation, feature frequencies, classification methods, data mining methods and neuro fuzzy method. We have only used rule generalization method for it. In future, we will be using any of these methods for reducing false alerts generated by Snort rules. *Second*, we have used static dataset “DARPA” for creating and testing our rules for different labeled attacks. In future, we will apply these rules on some dynamic dataset like BSNL server or some website data so that these rules can be made more efficient in all perspectives. *Third*, we will be using Snort++ or Snort 3.0 in future for further advancement in our rules to detect intrusions over the network. Snort++ alfa version is out now in the market for testing purpose.

REFERENCES

- [1] J. Kuang, L. Mei, and J. Bian, “An innovative implement in organizing complicated and massive intrusion detection rules of IDS,” in *2nd International Conference on Cloud Computing and Intelligent Systems*, Hangzhou, Vol. 03, pp. 1328–1332, 2012.
- [2] Ritu Makani, Yogesh Chaba “Analysis of Security Techniques for Computer Networks”, in *International Journal of Engineering Research in Computer Science and Engineering*, Vol. 1, pp. 1-3, 2014.
- [3] “SNORT Users Manual 2.9.8.2.” [Online]. Available: <http://manual-snort-org.s3-website-us-east-1.amazonaws.com/>.
- [4] M. Roesch and S. Telecommunications, “Snort - Lightweight Intrusion Detection for Networks”, in *Proceedings of USENIX LISA*, Seattle, Washington, USA, Vol. 99, No. 1, pp. 229–238, 1999.
- [5] E. Kostlan, (2015), “Intermediate - Snort Implementation in Cisco Products”, *CiscoLive 365*.[online]. Available: http://www.ciscolive.com/online/connect/sessionDetail.ww?SESSION_ID=83682.
- [6] “MIT Lincoln Laboratory: DARPA Intrusion Detection Evaluation.” [Online]. Available: <https://www.ll.mit.edu/ideval/data/2000data.html>. [Accessed: 09 Apr-2016].
- [7] G.D. Kurundkar, N.A. Naik and S.D. Khamitkar, “Network Intrusion Detection using SNORT”, in *International Journal Of Engineering Research and Application*, Vol. 2, Issue 2, pp. 1288-1296, 2012.
- [8] Riad, Alaa El-Din, Ibrahim Elhenawy, Ahmed Hassan, and Nancy Awadallah "Using Jquery with Snort to Visualize Intrusion.", in *International Journal of Computer Science*, Vol. 9, Issue 1, No. 3, pp. 486-491, 2012.
- [9] R. Hilden and K. Hatonen, “A Method for Deriving and Testing Malicious Behavior Detection Rules,” in *IEEE Trustcom/BigDataSE/ISPA*, Helsinki, Vol. 1, pp. 1337-1342, 2015.
- [10] S. Schmerl, H. Koenig, U. Flegel, M. Meier, and R. Rietz, “Systematic Signature Engineering by Re-use of Snort Signatures,” in *Annual Computer Security Applications Conference*, Anaheim, California, pp. 23–32, 2008.
- [11] N. Khamphakdee, N. Benjamas, and S. Saiyod, “Improving Intrusion Detection System based on Snort rules for network probe attack detection,” in *2nd International Conference on Information and Communication Technology*, pp. 69–74, 2014.
- [12] A. Kirk, “Using Snort Fast Patterns Wisely For Rules”, *Cisco Talos Blog* 2010.
- [13] S. Terry and B. J. Chow, “An assessment of the DARPA IDS evaluation dataset using Snort”, in *UCDAVIS department of Computer Science*, Vol. 1, pp. 22-41, 2007.
- [14] A. Saboor, M. Akhlaq, and B. Aslam, “Experimental evaluation of Snort against DDoS attacks under different hardware configurations”, in *2nd National Conference Information Assurance*, Rawalpindi, pp. 31–37, 2013.
- [15] “Sourcefire VRT: Focused on protecting your network”, *Sourcefire Whitepaper*, 2012



Ayushi Chahal received M.Tech degree in Computer Science and Engg. from Guru Jambheshwar University of Science and Technology, Hisar , Haryana , India in 2016. She has received her B.Tech degree in Computer Science and Engg. from

Bhagat Phool Singh Mahila Vishwavidyalaya, Sonipat, Haryana, India. Her research interest includes Network Security, Big Data, IoT.



Ritu Makani is a Associate Professor in Computer Science and Engg. having 14 years of teaching and research experience. She has received her PhD. degree in 2014 from Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India. She has done her M.Tech in 2002 from Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India. Her core area of interest in research is Network Security.

A Low Cost ECG Monitoring System with ECG Data Filtering

Md. Rakib Hasan

Dept. of CSE

Jahangirnagar University

Dhaka, Bangladesh.

riyadrakib@gmail.com

Mohammad Rabiul Alam Sarker

Dept. of CSE

Jahangirnagar University

Dhaka, Bangladesh.

rabiulalam.jucse@gmail.com

Md. Firoz-Ul-Amin

Dept. of CSE

Jahangirnagar University

Dhaka, Bangladesh.

real.firoz@gmail.com

Mohammad Zahidur Rahman

Dept. of CSE

Jahangirnagar University

Dhaka, Bangladesh.

rmzahid@junix.edu

Abstract— The ElectroCardioGram (ECG), a set of graphs of electrical heart activity, is the principle tool used in diagnosis of different heart conditions. This paper illustrates the design and implementation of a low-cost ECG monitor using microcontroller, Arduino programming language and Raw ECG data filtering with different digital filters. This paper describes the development of accurate monitoring of a heart rate based on a microcontroller. We can record the ECG signals and Heart beats of all patients in a single computer. These biomedical signals are acquired and then processed with a microcontroller. For the patient suffering from the cardiac disease it is very necessary to perform accurate and quick diagnosis. For this purpose a continuous monitoring of the ECG signal, patient's current heart rate and BP is essential. We can monitor the patient's ECG signal by using Arduino board and ECG shield and receive in the central place in any hospital. We use the C/C++ programming to retrieve the ECG data obtained from the human body. We filter the data with different digital filters to remove noise form the ECG data and store the data into a text file. Those data can be plotted to have ECG waveform to diagnosis heart problems.

Keywords- *Electrocardiogram, ECG signal, FIR Filters, High-pass Filter, Low-pass Filter, Microcontroller*

I. INTRODUCTION

The ECG monitoring system generally reflects the electrical activity of human body. The cardiovascular diseases are measured from the ECG based on the abnormality in the parameters of that graph. Different parameters denote different level of cardiac problems. It is necessary to pass the ECG data to a specialist person. As this data is so sensitive in patient's perspective, in wrong hand it can be devastating for them.

An electrocardiogram is a test that checks for problems with the electrical activity of heart. It checks the abnormality of heart. It results from diastole and systole phases of heart. Diastole and systole represent the resting or filling phase of a cardiac chamber and the contracting or pumping phase, respectively. The characteristics of the ECG signal, including the heart rate, the PR interval, the QRS duration, the QT interval, etc., are the important evidence for doctors to diagnose diseases. The change in these parameters indicate

illness of heart. If there happens any variation to process the ECG waveform it may cause misdiagnosis. So it is very essential to process the ECG waveform with a great care so that we get the real ECG waveform. After the collection of ECG data it is very much essential to filter those data because those data may contain noise and unwanted data. Thus, the target of ECG filtering is to reduce the redundancy as much as possible while to maintain clinically acceptable signal quality [1] [2].

Filtering is the process of removing unwanted data form a signals. It refers to removing noise and amplifying some data to get the actual data. Some common filters are used to remove the noise and unwanted data from ECG data. Low-pass filter, High-pass, FIR filter and QRS detection algorithm are used to have the actual ECG waveform form those actual ECG data are collected form human body through Microcontroller Based system. In this paper Low-pass, High-pass and FIR filters are used to remove the unwanted data and QRS detection algorithm is used for making the ECG waveform smoother.

II. LITERATURE REVIEW

ECG devices are used to asses heart rhythm (rate and regularity of heartbeats), measure sizes position of chambers, also the presence of any damage of heart and the effects of the heart. To diagnose poor blood flow to the heart muscle, heart attack ECG devices are used. Devices such as Pacemaker are used to regulate the heart.

Electrocardiography (ECG or EKG from Greek: kardia, meaning heart) is a transthoracic (across the thorax or chest) interpretation of the electrical activity of the heart over a period of time, as detected by electrodes attached to the surface of the skin and recorded by a device external to the body [3]. The recording produced by this noninvasive procedure is termed an electrocardiogram (also ECG or EKG). ECG devices are performed to diagnosis or research about human heart; they are also used on animals for research and non-nature research.

ECG data are useful for research. Storing the ECG data can be very handy for the further use. But raw ECG data may include

some noise, baseline wander. This causes complexity for ECG data usage. Applying different filters can eliminate those constraints.

Masaki Kyoso proposes a simple algorithm to detect abnormal ECG. The algorithm is composed of a baseline drift canceller (utilizing a moving average calculation), a waveform detector (using a modified second order derivative) and an ECG analyzer[4].

F. Buendía-Fuentes, M. A. Arnau-Vives, A. Arnau-Vives propose a High-Bandpass filtering technique for ECG data to remove baseline wander. They try to find out the error in the interpretation of ST segment [19].

S. Sundar, S. Karthick and S. Valarmathy implementss FIR filter with Canonical Signed Digit (CSD) to remove noise from ECG data. They present the study of FIR filter using common subexpression elimination techniques for ECG signal Processing [5].

III. SYSTEM IMPLEMENTATION AND DATA FILTERING

In this research we attempt to design a microcontroller based ECG embedded system, especially for diagnosis of heart related problems using some hardware toolkit for rural medical center of Bangladesh. For this proposed system here we have to use a hardware toolkit, which is capable of measuring daily health conditions of electrocardiogram (ECG) and this digital signal is transfer to a receiving device for signal processing and we using EKG shield for convert them as usable form of binary data. We just collect the ECG data and filter it using different ECG data filtering algorithm (like: Low-pass filter, High-pass filter, FIR filter). Then this raw data will be used for prescribe medicine by the cardiac specialist by the signal showing curve.

System Model: In this project we have tried to design, develop and implement a system that will create the initial step towards a low cost microcontroller based ECG measurement system which can be a great solution for the welfare of the rural people in terms of their cardiovascular diseases. During this phase of that proposed system we have sent data from human body to a computer situated at the rural health care center using Arduino ECG-EKG shield. In that purpose we have to use Arduino Programming Language (widely known as "Micro C") to create an interface between the microcontroller based ECG device and a USB communication port of that computer.

When ECG data started coming from the human body through the shield, we have written a program in C programming language to receive those data from the COM port and save them in a .txt for further use.

In that sense, we can divide the implementation procedure of our project into several phases. They are as following:

1. Hardware Setup
2. Interfacing between Arduino Shield and USB port of the computer
3. Reading serial data from the COM port
4. Storing those data into a text file
5. Filtering the data using ECG data filtering algorithms

The stored data from the human body can be used for various purposes in future. As per the proposed system the data will be sent to the doctor's smartphone or any other portable device which will create an ECG waveform in that device. From the created graph the specialist physician will be able to diagnosis the cardiovascular condition of the patient and prescribe instantly as well. As a result after implementation of this system, the distance between the rural patients and the doctors will be lessen and it will have a significant impact on the national healthcare condition.

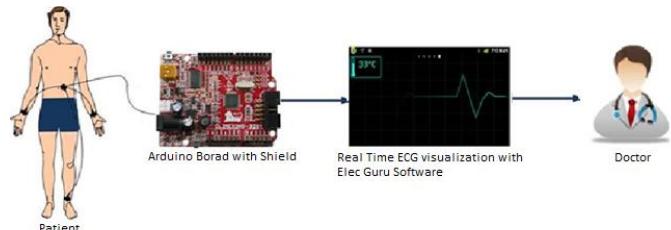


Figure 1: Real-time ECG measurement system

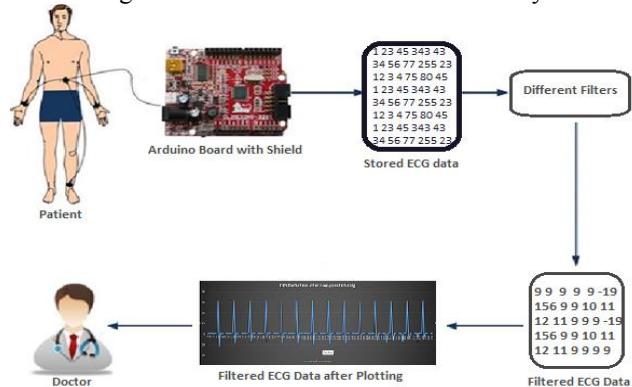


Figure 2: Filtering Raw ECG Data

Hardware Setup: Setting up the hardware equipment is a critical and sensitive step of the implementation procedure. We had to take this step so cautiously because if the board is exposed to high electrostatic potentials then any kind of permanent damage can be occurred. There are three Hardware equipment needed to implement this project.

1. Arduino Compatible Board (Ex. Olimexino-328)
2. Shield EKG-EMG (Ex. Olimax EKG Shield)
3. Electrode Cable (Ex. EKG-EMG PA)

Shield EKG-EMG is the major hardware product of this project. It is built by the Olimex group. It is a microcontroller based board which is being added to an Arduino compatible

board to make a device to measure the heart rate of human body.

For the purpose of this project, we have used Olimexino-328 Arduino compatible board. Olimaxino-328 is compatible both with Olimax EKG shield and EKG-EMG Pa electrode cable. Electrode cable is connected to the human body in order to response to the electric signal of the human body.

Electric Guru Software to visualize real-time ECG: We use Electric Guru Software toolkit to represent the pulses of heart as signal. Our heartbeats are rated into the signal by this software. Arduino IDE communicates with the software through the serial communication port. Binary data bits are generated according to the pulse signal. Every signal rounds up to one cycle and again start.

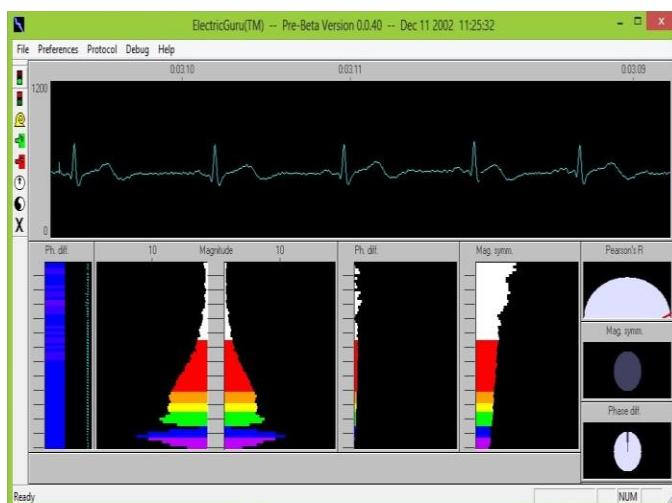


Figure 3: ElecGuru Software for real-time ECG visualization

Reading and Storing Data from Serial Port: We used C programming language in order to read the ECG data from the serial port. We wrote programs using the C Library for serial operation in order to fetch the ECG data. We also use file manipulation operations of C to store those data in a text file of local computer.

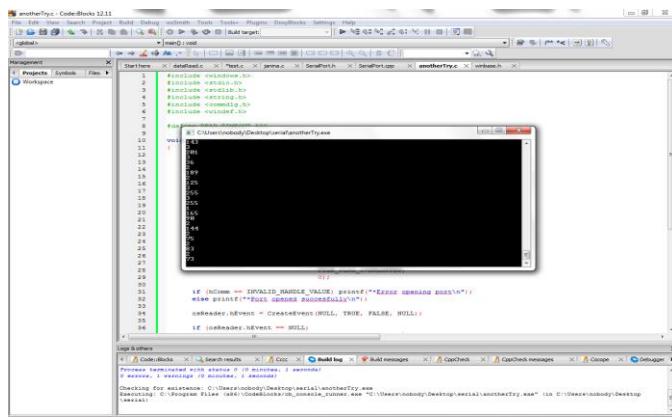


Figure 4: Reading and storing ECG data using C programming

Filtering the stored ECG data using digital filters:

Stored data may contain Baseline wander, Power line interference and muscle noise. To eliminate those noise we apply different filters including High-pass filter, FIR filter, Low-pass filter and QRS detection algorithms. Using of those filters certainly make the ECG data more practical and functional.

Low-Pass Filter: A Low-pass filter is a filter that passes signals with a frequency lower than a certain cutoff frequency and attenuates signals with frequencies higher than the cutoff frequency [6][10]. The amount of attenuation for each frequency depends on the filter design.

We use the Low-pass filter for limiting the ECG data sample and smoothing the ECG curve [8]. In this Low-pass ECG data filtering system we collect the ECG data from the Arduino and sample the data set with some parameter [9]. We remove the value from the data set that seems noise. We put the data set into a limited range. We implemented the Low-pass filter in C++ and filtered the ECG data.

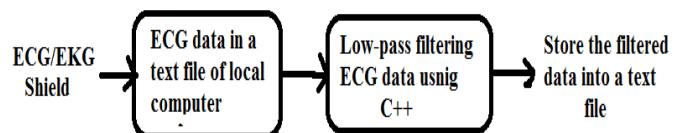


Figure 5: Implementing Low-Pass filter for ECG data

Low pass filter can mitigate the muscle noise.

High-pass Filter: Basically we implement a High-Pass filter to cut off the lower-frequency components such the baseline wander [20, 21].



Figure 6: Implementation of High-Pass filter for ECG data

Most biological signals must be processed for adequate recording. High-pass filters may distort the shape of the recorded signal and sometimes may cause electrocardiographic changes simulating myocardial ischemia. High-pass filters cut off the low frequency and let the high frequency pass by [19].

FIR Filter using Hanning window: A finite impulse response (FIR) filter has a unit impulse response that has a limited number of terms [12]. FIR filters are generally realized non recursively, which means that there is no feedback involved in computation of output data [11]. The output of the filter depends only the present and past input. The difference equation is:

$$Y(nT) = \sum_{k=0}^N b_k x(nT - kT)$$

And a transfer function of this filter is given below:

$$H(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_N z^{-N}$$

Most common simplest smoothing filter which reduces the high frequency noise is Hanning filter [16]. By difference equation the Hanning filter computes a weighted moving average. We have the difference equation representing the numerical algorithm for implementing a digital filter. Hanning filter equation is:

$$y(nT) = \frac{1}{4} [x(nT) + 2x(nT - T) + x(nT - 2T)]$$

Hannig filter can be implemented by writing a computer program in C++ language where data has previously sampled and store in a text file [17]. This program directly filter the digital ECG data.

QRS Detection Algorithm: QRS detection algorithm is a technique which helps to detect the QRS complex of the ECG signal. Preprocessing and decision are the two different stage of QRS detection algorithm [13]. In preprocessing step ECG data pass a block of filter to remove the noise. Than in decision step the examined or predefined ECG data compares filtered ECG data for detecting the QRS complex wave [14] [15].

Here we first go through the Low-pass filter in preprocessing step of QRS detection [16]. Then we compare the ECG data set with the collected data set means predefined ECG data for QRS complex detection in decision step. For QRS complex detection, here we take the baseline data as a “Q” data, peak value as “R” data and the lowest value as “S” data.

Now we also determine the QRS complex after FIR filtering of ECG data.

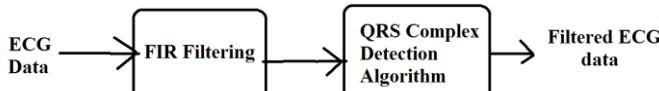


Figure 7: Implementing FIR filter for ECG data and Applying QRS detection

By this QRS detection we can filter any ECG data but we must have a standard ECG dataset to compare and to remove those abnormal data from the dataset [7]. It is very useful for getting the actual ECG data

Advantage of those filters:

Microcontroller based ECG measurement system is not noise free. There are some noise and unwanted data mixed in the dataset that we collect from the human body. So it is essential to remove those noise and unwanted data from the dataset to get the actual ECG data. The advantages of using those filters are given:

1. The Low-pass filter is used to remove the unwanted spike (data that is not peak value R) from the dataset
2. Applying QRS filtering algorithm to detect the QRS data from the dataset.

3. FIR filter is a digital filter which is more accurate than analog filter [17].
4. FIR filter has a linear phase characteristic so that it can remove the baseline wander and power line interference [18].
5. QRS detection algorithm is more important because the energy of the heart bit located on the QRS complex [13].

IV. RESULT ANALYSIS AND DISCUSSION

This project aims basically on extracting the data from human body which is being used to measure the ECG wave by Arduino EKG-EMG shield. In order to analyze the ECG data of human body from different ages, we have investigated this system on different people. We have got following result based on that investigation.

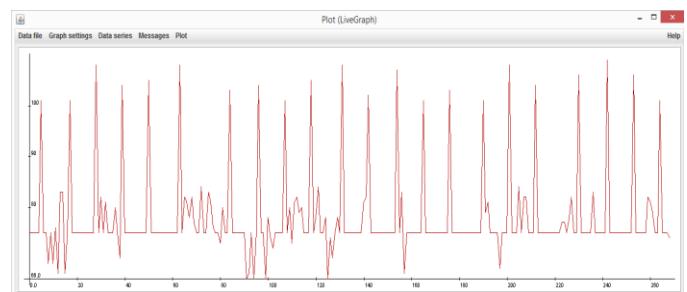


Figure 8: Raw ECG data plotted on Live Graph Software

From figure-8, we have seen that raw ECG data have noise and baseline wander. We found a minimum similarity with actual ECG graph. But after applying those filters on the same ECG data we got significant changes which is shown in figure-9. The obtained outcome is very close to actual ECG data and it became more reliable to know the cardiac health situation of patients.

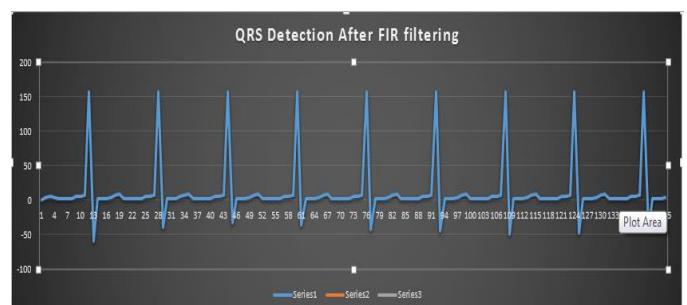


Figure 9: ECG graph after filtering and QRS detection

If we can find the number of errors, it can be a significant way to extract and store the heart rate data of human body. We tried to filter those ECG data using different filters for removing the noise form the ECG data. Further research can contribute to the improvement of this process and make it more accurate.

V. CONCLUSION

For rural medical center, a microcontroller based ECG system is distributed information system to connect rural patients to

cardiac specialists. Therefore, it is of significance to consider interoperability of ECG systems not only in rural people but also for the personal use. We mostly focus the project utilization in one's own diagnosis of cardiac problem by oneself. Home medical diagnosis encourages this ECG embedded system.

Available use of device will ensure the diagnosis of any type of cardiac problem by the communication to the cardiac specialist within a short time. Filtered ECG data can be very useful for the cardiac specialist to detect anomalies regarding cardiac problems. In future more filtering is needed to have the exact ECG waveform. These are some basic filters that are used to clarify ECG data from noise and baseline wander. This system needs more research for getting exact ECG waveform.

REFERENCES

- [1] Digitally Filtered ECG Signal Using Low-Cost Microcontroller, Asiya M. Al-Busaidi and Lazhar Khriji, Dept. of Electrical and Computer Engineering , Sultan Qaboos University, Muscat, Oman
- [2] Boston In book: Clinical Methods: The History, Physical, and Laboratory Examinations, Edition: 3rd, Chapter: Chapter 33, Publisher: Butterworths, Editors: H Kenneth Walker, W Dallas Hall, J Willis Hurst Source: PubMed
- [3] Kumar, Ashwini. ECG-simplified. s.l. : Life Hugger, 2011.
- [4] Masaki Kyoso, A Study of Data Reduction Method for ECG Medical Telemetry System, IEEJ Transactions on Electronics, Information and Systems, Volume 122 (2002) Issue 9 Pages 1595-1602
- [5] S. Sundar, S. Karthick and S. Valarmathy, Filtering Noise from Electrocardiogram using FIR filter with CSD Coefficients, International Journal of Computer Applications
- [6] Soderstrand, M. A. 1972. On-line digital filtering using the PDP-8 or PDP-12. Computers in the Neurophysiology Laboratory, 1: 31-49. Maynard, MA: Digital Equipment Corporation.
- [7] Forno, G. S. and Tompkins, W. J. 1982. QRS detection using automata theory in a batterypoweredmicroprocessor system. IEEE Frontiers of Engineering in Health Care, 4: 155–58.
- [8] Abenstein, J. P. and Tompkins, W. J. 1982. “New data-reduction algorithm for real-time ECG analysis”, IEEE Trans. Biomed. Eng., BME-29: 43–48
- [9] Oppenheim, A. V. and Willsky, A. S. 1983. “Signals and Systems. Englewood Cliffs”, NJ: Prentice Hall.
- [10] Prakash Vidwan and V.T Patel, “Real Time Portable Wireless ECG Monitoring System”. Asian Journal Of Computer Science And Information Technology 2: 6 (2012) 158 – 161.
- [11] Rabiner, L. R. and Rader, C. M. 1972. Digital Signal Processing New York: IEEE Press.
- [12] Antoniou, A. 1979. Digital Filters: Analysis and Design. New York: McGraw-Hill
- [13] Raul Alonso Alvarz, Arturo J. Mendez Penin and X. Anton Vila Sobrino. “A comparison of three QRS detection algorithms over a public database”, Procedia Technology 9(2013) 1159-1165, HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.
- [14] J. Pan and W. J. Tompkins, “A real-time QRS detection algorithm”, Biomedical Engineering, IEEE Transactions on, vol. BME-32,1985.
- [15] Bert-Uwe Kohler, Carsten Henning and Reinhold Orlmeister, “The Principles of Software QRS Detection. Reviewing and Comparing Algorithms for Detecting this Important ECG Waveform”, IEEE ENGINEERING IN MEDICINE AND BIOLOGY January/February 2002.
- [16] W Zong, GB Moody and D Jiang, “A Robust Open-source Algorithm to Detect Onset and Duration of QRS Complexes”, Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA.
- [17] Bhumika Chandrakar, O.P.Yadav and V.K.Chandra, “A SURVEY OF NOISE REMOVAL TECHNIQUES FOR ECG SIGNALS”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 3, March 2013.
- [18] J. A. Van Alste, T. S. Schilder “Removal of Base-Line Wander and Power-Line Interference from the ECG by an Efficient FIR Filter with a Reduced Number of Taps” IEEE Transactions On Biomedical Engineering, vol. bme-32, no. 12, December 1985 page no-1052-1060
- [19] F. Buendia-Fuentes, M. A. Arnau-Vives, A. Arnau-Vives, Y. Jiménez-Jiménez, J. Rueda-Soriano, E. Zorio-Grima, A. Osa-Sáez, L. V. Martínez-Dolz, L. Almenar-Bonet, and M. A. Palencia-Pérez, High-Bandpass Filters in Electrocardiography: Source of Error in the Interpretation of the ST Segment, ISRN Cardiology, Volume 2012 (2012), Article ID 706217, 10 pages
- [20] Christov, I., I. Dotsinsky, I. Daskalov, High-pass filtering of ECG signals using QRS elimination, Med. Biol. Eng. Comp., Vol. 30, pp 253-256, 1992.
- [21] Ivo Tsvetanov Iliev , Serafim Dimitrov Tabakov , Vessela Tzvetanova Krasteva, COMBINED HIGH-PASS AND POWER-LINE INTERFERENCE REJECTER FILTER FOR ECG SIGNAL PROCESSING, Dec 2010 International Journal Bioautomotion

AUTHORS PROFILE

Md. Rakib Hasan: He Completed his graduation in CSE from Jahangirnagar University and Masters in CSE from Jahangirnagar University.

He has 2 years of teaching experience. His research interests are Machine Learning, Natural Language Processing, Network Security, Telemedicine and IoMT.

Mohammad Rabiul Alam Sarker: He Completed his graduation in CSE from Jahangirnagar University and Masters in CSE from Jahangirnagar University.

He has 2 years experiences as programmer in BJIT limited. His research interests are Machine Learning, Telemedicine, Natural Language Processing.

Md. Firoz-Ul-Amin: He Completed his graduation in CSE from Jahangirnagar University and Masters in CSE from Jahangirnagar University.

Now he is persuing his PhD from Louisiana State University. His research interest includes E-Commerce, Telemedicine and IoMT.

Mohammad Zahidur Rahman: He Completed his graduation from BUET, Masters from BUET and PhD(Malaysia).

He has about 20 years of teaching experience. His research interests are E-Commerce, Computer Security, E-Governance, Communication , Telemedicine and IoMT

The convenience activity on TQM of Advanced Technology on user expectation of online Banking Systems in India

Mohd Faisal Khan¹,
IPhD Scholar,
Department of Information Technology (IT),
AMET University Chennai,
Tamil Nadu (India),
M_faisalcse@rediffmail.com

Dr. Debaprayag Chaudhuri²
Approved Ph.D. External Guide
Department of Information Technology (IT),
AMET University Chennai
Tamil Nadu (India)
debaprayag@gmail.com

Abstract (Size 10 & bold & Italic)— This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IJCTT Journal. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

Abstract- Indian growing new financial establishments, notably banks, square measure one among the most important investors within the domains of knowledge systems, and there square measure quite clear signs that these trends to spread within the future. The arrival and enlargement of globalization and also the development of Advanced technologies knowledge pushed the banks to adopt advanced technology to launch new services. Banks have applied remote enabled service victimization the net to achieve competitive advantage, increase potency, scale back prices and provide a range of latest services. On-line systems create banking transactions straightforward and convenient, notably for disabled people that could need special services. the most purpose of this current study is to look at the most keys to live the advantage perception of victimization net banking technology, as this advanced technology is taken into account together of the principal motivations underlying the inclinations of people to adopt such a convenient technology in India. The model developed associated developed during this analysis study is an extension to the Technology Acceptance Model (TAM). The model was tested with a survey sample of four hundred folks chosen arbitrarily. The findings of the study indicate that everyone mentioned factors within the projected model (CNV, SE, QI, AW, PEU, PU) have important impact at intervals prompting the employment of net banking systems. The information analysis relies on the applied mathematics Package for scientific discipline (SPSS).

Keywords: Online banking; TAM; CNV, SE, QI, AW, PEU, PU, TQM

I. INTRODUCTION

Over the past few decades, the world economy is undergoing Associate in extraordinary growth of Information & Communication Technology that has affected the entire life. The development of the ICT and digitalization have opened a new window of communication for people and businesses and provided opportunities to speak and acquire information in a completely different approach. Advanced ICT has become a very important issue for each global economy and its related segments. The service segments, significantly the banking sectors, are continually growing; customer's area components more and more unpredictable and facility distribution may be a turning opinion. Since to the growing implement through the ICT and developing advanced technology systems globally, there has been a ostensible intensification within the usage of e-banking, online banking, tale banking etc adopting through the world.

The progressively competitive setting within the monetary services sector besides globalization, alleviation, and advanced technology revolution has opened the door for brand new economical delivery channels also as additional innovative product and services within the industry. for example, web banking services supply a spread of advantages conduct online transactions faster and additional simply with self-service applications in terms of transfers between accounts, pay bills to utility suppliers and web buying. This additionally reduces operational prices for banking. for example, face-to-face dealings with a person's teller value and also the got to print receipts is significantly quite internet dealing.

According to Hoehle¹ instructed that each one analysis associated with e-banking encompasses varied disciplines of selling, e-commerce, system, business and management. world net users square measure progressively disbursal longer on-line. Owing to this, the banks in most countries offer their services on-line to stay their on-line customers. This helps those users to perform most of their banking transactions solely by visiting the bank's web site,

and while not being physically gift within the bank. The online banking accessibility is generally helpful in terms of making certain people square measure able to access the web content; particularly for disabled individuals WTO might need special services. The term of incapacity here is outlined because the consequence of physical impairment that leads to restrictions on the flexibility of a human movement in society. A incapacity could also be gift from birth, or occur throughout a personality's period. Therefore, net banking systems return exactly to serve all classes of people, significantly in terms of overcoming the disabling physical negative effects.

This ensures that the net banking may be a tool that enhances everyone's ability to access info, instead of a tool of exclusion. making accessible content ought to be associate integral a part of developing a bank's computing machine, and a thought of accessibility necessities ought to be incorporated into all aspects of the look method. most significantly, this can facilitate promote a additional inclusive digital world wherever resources is shared and employed by each individual.

Furthermore, some lecturers have centered on client self-service technologies, highlight the importance of technology used as a service enabler for the client²⁻⁵. the advantages of such technologies square measure argued to stem from the very fact that customers will access services once and wherever they need while not a number of the complications of social exchanges Bitner³. web banking industry is one amongst such technologies, and forms the final study of this current study.

II. OVERVIEW OF INTERNET BANKING TECHNOLOGY

The Internet is actually a world development, creating each of distance and time unsuitable to several exchanges. Internet industry assures the exploitation of latest business opportunities within the banking sector in terms of simpler performance, larger economic potency, and a faster exchange among money markets. Globalisation of advance technology and demographic a major trends poignant the economy in every country Seipp⁶. This suggestion increasing competition among banks and different money establishments. In addition, the internet and advanced ICT have brought radical changes to the banking sector in step with Ody⁷, individuals use the internet for the most part for 2 main reasons; to seek out info or obtain merchandise and repair handily in faster pace. He conjointly emphasizes the importance of quality characteristics on system acceptance. This inspired most banks to produce a spread of on-line services, which permit their customers to perform most of their banking transactions by visiting the bank's web site.

Internet banking industry is associate innovative style of advanced information systems technology designed for end- users and offers them on-line services that change them to conduct their money transactions through a laptop devices or pill devices recently. Unless the user has personal device and convenience of the net, it's unlikely to contemplate victimization online banking in any respect⁸. Banking on-line system includes digital series of processes, whereby shoppers square measure able to log into the bank's web site through the web-browser put in on the laptop and perform varied on-line transactions by employing a personal username and positive identification supported the user's choice. Additionally, web banking systems support communication with alternative servers, like web info servers. These participate within the setting and contribute alternative services and data to gift a range of on-line services. Banks produce their formal computing device through the adoption of basic internet technologies. Figure 1, clarifies the most useful parts, their roles and contribution among the complete system, whereby the end-user will access the secure web site of the bank via the net. Interaction between user and repair supplier systems is supported by structure dialogs. The checking account server as a part of the system receives the directions to supply the substantial functions to be performed on the bank accounts, whereby finish users assume a lot of responsibility for his or her own applications, and that they seldom have direct interaction with the operations employees of the websites.

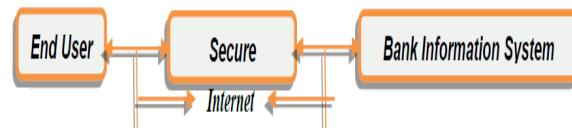


Figure 1: Relationship between users and banking server.

The developments in ICT have had a huge result within the development of additional versatile payment strategies and additional easy banking services⁹. The event and diffusion of web banking technology area unit expected to lead to additional economical banking systems. Additionally, banking establishments can give their product and services through such electronic channels, additional handily and economically while not reducing the standard of the prevailing levels of service on blessings of online banking systems area unit various for each banks and customers. For banks, online banking brings a variety of advantages from reducing prices and Time to realize bigger satisfaction for his or her customers. For customer, it provides them a simple access to their accounts, as they not have to be compelled to visit banks to try and do their transactions in person. Service suppliers additionally get pleasure from web banking because it thought to be the simplest means of achieving growth. to boot, web banking provides

another for quicker delivery of banking services compared to the standard strategies.

According to the newest statistics, the amount of net users round the world in 2016 is calculable at concerning three.7 billion, a rise of over 933.8 % between 2000 and 2016¹⁰. This result affirms that internet users are massively increasing across world countries, that creates the online banking systems represents the biggest transactional sector on the online. While, India is taken into account the guts of geographical region countries that drove most Indian banks to adopt on-line banking services so as to satisfy the requirement of their customers as a results of last a trade Agreement between India and a few foreign countries and organization like WTO, World Bank, G20 group, G8 Group, the U.S, whereby the entire variety of internet users in India from 2015 to 2022. In 2017 India had 331.77 million internet user. This figure projected to grow to 511.89 million internet user in 2019 is calculated by Indian business forum at Feb 2018¹¹. This result quiet clarifies that telecommunication and net sector is one in every of the quickest growing industries in India.

Despite of studies that explore the convenience of online banking technology are gettable, analysis within the context of Indian perspective remains insufficient. Therefore, the most objective of this study is to grasp the substantial factors to live the extent of advantage perception of victimization such a convenient technology by developing a digitalization economy (for community in India), that raises their intentions towards the employment of online banking. To realize the analysis objective, one main analysis question was addressed: "What factors have an effect on the advantage perception of victimization the net banking technology from the user's perception?".

III EMERGENCE OF RESEARCH MODEL

The initial adoption of on-line services like web banking, primarily involves the acceptance of each the web technology and on-line service suppliers. Many approaches were developed so as to look at and perceive the factors moving the acceptance of technology in organizations, as well as the idea of reasoned action (TRA)¹², the idea of planned behaviour (TPB)¹³, The model of laptop utilization¹⁴, The rotten theory of planned behavior^{15,16}, innovation diffusion theory (Rogers, 1983, 1993; Agarwal and Prasad, 1997), and therefore the moguls model of computing¹⁷. Withal, the technology acceptance model (TAM) is basically employed by specialist researchers within the domain of data systems thanks to its quality with high validity. During this context, (TAM) is applied during this current study as a theoretical background for a few reasons:

- It's the foremost effective model within the field of data systems and technology for testing user acceptance and usage behavior¹⁸,

- It's a prognostic power that makes it straightforward to use in numerous things¹⁹,
- There's a standard agreement among researchers that the model is helpful in predicting individual's acceptance of varied technologies^{20,21},
- It helps to know the connection between totally different instructive variables.

Technology acceptance model (TAM) is principally steered for technology-based perspective through dual system options of perceived utility (PU) and perceived easy use (PEU). Perceived utility is outlined because the extent to that someone believes that mistreatment specific technology would enhance her/his job performance whereas perceived easy use is that the degree to that mistreatment it's freed from effort for the user²². The model distinguishes Perceived utility (PU) and Perceived easy Use (PEU) as key factors that influence acceptance of a precise technologies. Within the gift study, the scientist assigns element within the context of web banking because the degree to that a user believes that mistreatment web banking industry service would enhance banking services usability. While, PEU is set because the degree to that a user believes that mistreatment web banking technology would be free from effort.

According to Davis²², analysis in technology acceptance should be self-addressed, however different variables have an effect on utility, easy use. various studies have sought-after to expand the cap by incorporating further constructs¹⁵. In accordance with previous studies, the abstract framework of this study is developed supported a review of the literature and changed by the author to create it relevant to the Indian scenario. As a result, associate extended of the cap model contains external variables will be employed in order to explore influential factors of build web banking systems a lot of convenient throughout the mode of usage. Thus, many hypotheses are shaped for investigation the theoretical model in India.

IV. EXTERNAL VARIABLES

Convenience (CNV): The online wordbook defines convenience as "anything that adds to one's comfort or saves work; helpful, handy or useful device, article, service, etc." within the selling context, is stated convenience product as those who the buyer purchases oft and forthwith at simply accessible stores Copeland²³. It should even be outlined as client perceptions concerning the relative time and energy gone in either getting or employing a service²⁴. Convenience has been one amongst the principal motivations underlying client inclinations to adopt on-line getting²⁵⁻³². Within the current study, the author defines convenience within the context of e banking as an automatic accessible on-line service twenty four hours every day and 7 days, that will increase comfort for users whereas reducing the expenditure of your time and energy on the a part of exploitation such a sophisticated technology.

The construct of service convenience is flat in nature^{24, 30, 33-35}. Many authors have acknowledged that service convenience impact on overall shopper assessment of the service, together with satisfaction with the service additionally as perceived quality^{24,36}. Moreover, Seders et al.³⁴ have extensively reviewed the literature on shopper convenience in an exceedingly service economy and outline “service convenience” as consumers’ time and energy perceptions associated with shopping for or employing a service. The time-saving side of convenience has been intensively investigated in shopper waiting literature, notably with relevance shopper reaction to waiting time³⁶. The idea of effort saving refers to the reduction of psychological feature, physical, and emotional activities that customers should bear to buy merchandise and services²⁴. For example, The need of consumers to get convenience and time-saving ability to look at and pay multiple bills in an exceedingly single place. In addition, the opposite dimension of convenience is that the accessibility term, whereby the ability of the net is in its accessibility by everybody. Accessibility term determines because the ability of users to access info and services from the web site, that essentially admit the content format; the user's hardware, software package and settings; internet connections; the environmental conditions and also the user's skills and disabilities. In line with internet banking industry, The accessibility of bank web site converges to the implementation of website content during a manner to maximize the power of various classes of people to access it. Karahanna and Straub additional urged that examine the impact of accessibility on the perceived easy use. Their analysis results indicated that perceived accessibility considerably and absolutely influences the construct perceived easy use. In the end, the author during this study focuses on the size of on-line convenience square measure in terms of access, search, evaluation, and group action.

H1: Convenience (CNV) features a positive impact on customer's perceived simple use.

Technology self-efficacy (SE): The technology self-efficacy is associate degree individual's belief regarding his/her ability to with success use the technological service to accomplish a particular task - a confidence no inheritable from multiple positive experiences and bought familiarity with the web channel. Self-efficacy construct has been examined within the data systems literature^{38, 39}. A study of Davis et al.²² prompt that the technology self-efficacy and therefore the construct of ‘perceived simple use’ area unit connected.

H2: The technology self-efficacy (SE) features a positive impact on customer's perceived simple use.

Quality of the internet association (QI): The standard of the internet association could be a major ingredient for any web-based applications. With improper web association, the employment of web

banking becomes not possible. Thus, confirms that there's a big relationship between the speed of the net and also the use of web banking services.

H3: Perceived quality of the net association (QI) contains a positive impact on customer's perceived quality.

Awareness of services (AW): Gaining awareness is largely authoritative in mistreatment web banking services which this should be achieved properly. In step with Sathy⁴⁰ and Al-Somali et al.⁴¹, awareness of service has direct influence on user intention to use the technology.

H4: Awareness of on-line services (AW) and its advantages contains a positive impact on customer's perceived quality.

V INTERNAL VARIABLES

Perceived easy use (PEU): This study suggests that the web banking industry needs less effort to use, learn, and train. AN empirical study conducted by Wang et al.⁴² shows that perceived easy use features a direct important positive impact on behavioural intention to use web banking. Geffen et al.⁴³ Any illustrate that perceived easy use, trust, and perceived quality are thought-about as important determinants of on-line searching. supported cap, an on the spot positive relationship has existed between PEU and element, which ends up in improved performance by saving effort required to try a similar work, that is proved by the rise in PEU^{20, 22, 44}. Within the lightweight of on top of context, PEU has AN influence on user acceptance of a web banking industry, each directly and indirectly through its impact on the element.

H5a: Perceived easy use features a positive impact on the intention to use the web banking industry.

H5b: Perceived easy use features a positive impact on the user's perceived utility of the net industry.

Perceived utility (PU): (PU) is one among the foremost standard and vital factors within the existing literature of on-line industry⁴⁵. This importance of technology element suggests that users area unit typically a lot of probably to simply accept a system primarily attributable to the functions it performs, implying that the benefit of use cannot catch up on a system that doesn't offer the desired practicality²². A lot of studies showed that perceived utility influence client interactions with web banking⁴¹, and these studies additionally advised that perceived utility affects the adoption of web banking services.

H6: Perceived utility features a important positive impact on the intention to use the net industry handily.

A regard to previous researches conducted with tam-o'-shanter in predicting new acceptance technology, actual usage is usually measured through activity intention (BI)^{13,46}. Therefore, the scientist here is set to travel in conjunction with previous studies and thought of intention to use (IU) because the variable

quantity rather than actual usage, for the essential reason that within the original tam-o'-shanter, chemical element and PEU were postulated to possess an on the spot relationship with the construct of intention to use however not with actual use. Figure 2, next exemplifies the hypothesized projected model.

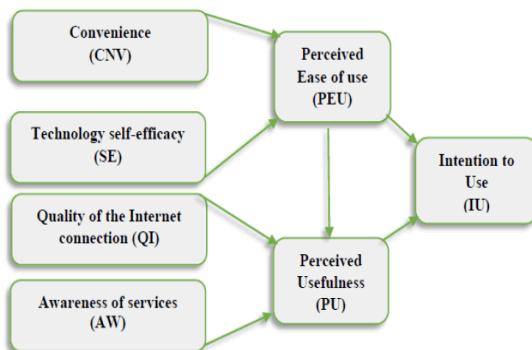


Figure 2: The proposed model.

VI EXPERIMENT DESIGN

The form could be a mechanism of information assortment, that is taken into account very fashionable among researchers. A self-administered form was created and developed supported previous literatures to get the most effective outcomes. Interval scaling within the variety of a numerical scale was elite because the most acceptable to live all variables of the study. During this scale, standards} allotted to point order and measure distance in units of equal intervals⁴⁷.

The Likert scale is accepted and treated as yielding interval knowledge by most of researchers⁴⁸. By employing a survey methodology within the current study, five-point Likert scales with ranged from "strongly disagree" to "strongly agree" were accustomed examine participants' responses for the most factors that fashioned the planned model. What is more, the questionnaires were two-handed out on to the sample and were collected back when a selected time to make sure the validity, accuracy, and also the credibility of the information.

The present study largely uses the closed-ended queries within the survey form to stay the context of the question same for all respondents⁴⁹. This additionally helps in reducing researcher's bias. The form is split into 3 main sections per web banking service usage. The election of the form things was derived from previous literature and data systems studies, that is taken into account the most supply of data in developing the analysis model and form.

The form begins with general (demographic) section incorporates queries that collect data regarding gender, age, level of education, and income. within the second section, the participant's area unit asked to produce background data on web usage. The third set of queries belongs to things of various constructs within the planned model to live the study variables. additionally, the survey form is escorted with a

canopy letter orthography the aim of this study to make sure confidentiality and privacy of the info assortment method, and to make sure the respondents to understand with whom they're dealing⁵⁰.

For this study, that targeted on the banking system at intervals technological frameworks, a sample of four hundred participants was chosen from the Indian community. All participants were bank customers elect haphazardly from corporations, universities, and totally different establishments and area unit alleged to have some expertise in mistreatment the online. The expected age of adult participants is eighteen years or older. during this current study, 356 questionnaires were came back out of four hundred distributed, representing a response rate of eighty nine of the pristine sample. once screening the questionnaires, Bastille Day of the forms (50 responses) were exempted from the analysis thanks to incomplete answers for many sections within the questionnaire. 300 and 6 usable responses were employed in the analysis, yielding a response rate of 86.

Findings and Discussion

The data model was refined through validation of the hypothesized structure model mistreatment applied math strategies. consequently, the projected hypotheses were tested at intervals a survey involving 306 banking customers residing in Indian. Collected quantitative knowledge were primarily analysed mistreatment applied math Package for Social Sciences (SPSS).

To meet the needs of this study, variety of applied math techniques is applied to check and interpret the results of the information analysis, as well as descriptive statistics, responsibility take a look at. for example, the descriptive statistics of the respondents' demographic characteristics were analysed and given initial in Table one shown below and outline of different hypothesis analysis square measure given next.

Demographic Characteristics Analysis

Table 1: Demographic characteristics of survey respondents.

Items	Categories	Frequency	Percent
Gender	Male	170	55.6
	Female	136	44.4
	Total	306	100
Age	18-25	132	43.1
	26-35	89	29.1
	36-45	54	17.7
	Above 45	31	10.1
	Total	306	100
Education	High school	35	11.4
	Bachelor degree	196	64.1
	Master degree	56	18.3
	Doctoral degree	19	06.2
Income	Total	306	100
	Less than 500 INR	109	35.6
Have you used the Internet before?	Yes	306	100
	No	-	-
	Total	306	100
How many years you have been using the Internet?	<1	-	-
	1-4	19	06.2
	=5	287	93.8

	Total	306	100
Where do you use the Internet from?	At home	187	61.1
	At workplace	101	33.0
	At university	18	05.9
	Total	306	100
Does your bank offer the online banking services?	Yes	306	100
	No	-	-
	Total	306	100
Are you using the online banking system?	Yes	306	100
	No	-	-
	Total	306	100

The results of participants' demographic characteristics demonstrate that largest proportions (55.6%) were male, and (44.4%) were females, but eventually every genders unit of measurement pattern infobahn banking technology at shut proportions. The foremost necessary proportion (43.1%) of respondents by cohort, were those inside the 18-25 years recent category.

However, kids represent the dominant socio-economic class (72.2%). Additionally, The survey respondents were sometimes well educated with over 64.1% holding bachelor degree and 24.5% having postgraduate qualifications. Currently out that everyone respondents have associate education level enough to produce correct answers to the shape. Supported the gain, the foremost necessary proportion (43.5%) of respondents, were those earning 500-1000 INR monthly.

The results collectively reveal that everyone subsamples use infobahn service. This result's not stunning, as in line with infobahn penetration among the entire population of India as of the highest of 2014 (86.1%). To boot, the foremost necessary proportion (93.8%) of respondents has been pattern Infobahn for 5 years or plenty of, considerably reception with large proportions reached to (61.1%). Moreover, all the respondents inside the survey answered 'yes' once asked if they are pattern internet banking system (Table 1).

A. Reliability Test

The dependability takes a look at of lives is assessed by examining the consistency of the respondents' answers to all or any things within the measure⁵¹. All of the measures utilized in this study show AN adequate dependability with Cronbach's alpha

values move between 0.72 and 0.89, as shown in Table a pair of. Besides, all prices were on top of the suggested value (>0.7), indicating robust validity and content consistency inside the queries for every construct in measurement relationships inside the hypothesized model. In different words, this finding demonstrates that everyone the factors utilized in this study square measure well-designed underneath the conditions of this survey.

Table 2: dependability take a look at.

No.	Constructs	Alpha
1	CNV	0.77
2	SE	0.72
3	QI	0.87
4	AW	0.79
5	PEU	0.83
6	PU	0.89
7	IU	0.87

VII. SIGNIFICANCE ANALYSIS OF RESEARCH HYPOTHESES

To assess the applied math significance of the analysis model, it's obligatory to think about the multivariate analysis price, that is employed to research the variations among the cluster of means that and their associated procedures. Table three reports results of research.

Table 3: Analysis of variables.

Hypothesizes path	Sum of squares	Asymp. Std. Error	F Value	Sig
H1: CNV → PEU	6.825	0.013	1.584	0.001
H2: SE → PEU	9.365	0.068	4.033	0.000
H3: QI → PU	8.123	0.027	2.204	0.008
H4: AW → PU	7.989	0.070	2.452	0.012
H5a: PEU → IU	11.357	0.035	5.195	0.000
H5b: PEU → PU	10.485	0.030	4.579	0.010

Hypothesizes path	Sum of squares	Asymp. Std. Error	F Value	Sig
H6: PU → IU	12.697	0.007	5.579	0.000

According to table 3 all, the on top of hypothesizes (H1-H6) results area unit statistically important at the amount of significance ($\alpha \leq \text{zero.05}$). This confirms that every one the analysis hypothesizes area unit completely confirmed and area unit connected.

Correlation Analysis of Variables

Table 4: Correlation analysis of variables.

Constructs	QI	AW	SE	CNV	PU	PEU	IU
QI	1	0.621**	0.551**	0.624**	0.725**	0.629**	0.653**
AW	0.621**	1	0.668**	0.565**	0.739**	0.613**	0.632**
SE	0.551**	0.668**	1	0.739**	0.685**	0.652**	0.695**
CNV	0.624**	0.565**	0.739**	1	0.778**	0.723**	0.712**
PU	0.725**	0.739**	0.685**	0.778**	1	0.781**	0.756**
PEU	0.629**	0.613**	0.652**	0.723**	0.781**	1	0.735**
IU	0.653**	0.632**	0.695**	0.712**	0.756**	0.735**	1

Table 4: Correlation analysis of variables.

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

Pearson correlations were calculated to spot the correlations between all latent variables and the way extent is said to every alternative. Finding of the quantity Pearson's correlations is listed in Table four. The correlations between multi latent constructs area unit moderately positive related and statistically vital at p-value $< \text{zero.01}$. Briefly, the findings of a correlation check show a support for projected hypotheses.

**Correlation is important at the zero.01 level (2-tailed).

Given that multiple regression between latent variables might need a tiny low however vital impact on the bias of path coefficients⁵², the author checked for potential multiple regression among freelance variables. A collinearity check discovered lowest collinearity with the variance inflation issue (VIF) of all constructs move between one.089 and 2.784. As a rule of thumb, it's most frequently suggested that the VIF worth ought to be less than ten. Table 5 summarizes the results of hypothesizes that shows that usually users in Jordan have well awareness of victimisation web banking services⁵³⁻⁵⁶.

Table 5: Summary of testing hypotheses.

Hypothesizes path	Reliability test	Correlation test	Approved
H1: CNV → PEU	✓	✓	✓
H2: SE → PEU	✓	✓	✓
H3: QI → PU	✓	✓	✓
H4: AW → PU	✓	✓	✓
H5a: PEU → IU	✓	✓	✓
H5b: PEU → PU	✓	✓	✓
H6: PU → IU	✓	✓	✓

VII. CONCLUSION

With the advancement of the Online banking, and advanced technologies, online customers will gain unlimited access to online banking services they have and luxuriate in a wider vary of selections in choosing services with extremely competitive quality. Therefore, sustaining a high level of online banking convenience has progressively become a key propulsion for patrons, with the aim of enhancing their loyalty to use such a sophisticated technology.

Using data systems in Indian banks appears to be very important to the success of today's banking systems. The present analysis study evaluates the extended and changed cap model, and examines the most keys to live the advantage perception of mistreatment the net banking technology apply in India, furthermore as influencing users' intentions to use such advanced technology. The study finds that on-line banking customers typically have a totally conscious of such services that are provided over the internet in India.

Obtained results of the analysis during this study approve that everyone mentioned factors within the projected model have common positive impact inside prompting the employment of advanced technology, i.e. internet industry, however with completely different degrees of influence on the internet customer's inclination. The study additionally illustrates that the construct of perceived quality is that the most potent issue on the advantage perception of mistreatment the net banking technology among the complete variables.

In a context inside that banking establishments area unit more and more committed to introduce a bigger rationality within the evolving operation processes,

and improve the service quality standards towards their customers. Banks managements ought to focus additional in developing economical action plans and techniques to fulfil the requirements of their on-line customers, and understanding the importance play role of data systems as important tools operate to boost banking online services as a competitive advantage and enhance their structure efficiency; like alter customers to operate additional severally and ability to conduct varied on-line transactions on their own. Eventually, this study contributes a stronger summary of understanding of advanced technology use in current measurements of services quality inside the banking sectors, particularly web banking systems in developing economies analogous to India.

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IJCTT LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IJCTT tran.cls and IJCTT tran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: "original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com)".

REFERENCES (SIZE 10 & BOLD)

- [1] Hoehle H, Scornavacca E, Huff S (2012) Three decades of research on consumer adoption and utilization of electronic banking channels: A literature analysis. Decision Support Systems.
- [2] Gwinner KP, Gremler DD, Bitner MJ (1998) Relational benefits in service industries: the customer's perspective. Journal of the Academy of Marketing Science.
- [3] Bitner MJ, Brown SW, Meuter ML (2000) Technology infusion in service encounters. Journal of the Academy of Marketing Science.
- [4] Selnes F, Hansen H (2001) The potential hazard of self-service in developing customer loyalty. Journal of Service Research.
- [5] Dabholkar PA, Bagozzi RP (2002) An attitudinal model of technology-based self-service: moderating effects of consumer traits and situational factors. Journal of the Academy of Marketing Science.
- [6] Seipp M (2000) Exploring emerging market models in the financial industry, part of the seminar in information management: advanced topics. European Business School.
- [7] Ody P (2000) The challenging task of building strong e-loyalty: customer relationship marketing. The Financial Times, p.16.
- [8] Khrais LT (2013) The effectiveness of E-banking environment in customer life service an empirical study (Poland).
- [9] Akinci S, Aksoy S, Atilgan E (2004) Adoption of internet banking among sophisticated consumer segments in an advanced developing country. International Journal of Bank Marketing.

- [10] Internet World Stats (2015) World internet usage and population statistics.www.internet-worldstats.com/stats.htm
- [11] Internet Indian stats (2018). Indian internet user www.statista.com/statistics/255146/number-of-internet-user-in-India.
- [12] Ajzen I, Fishbein M (1980) Understanding attitudes and predicting social behavior. Prentice-Hall, Englewood Cliffs, NJ, USA.
- [13] Mathieson K (1991) Predicting user intentions: Comparing the technology acceptance model with the theory of planned behaviour. *Information, Systems Research*.
- [14] Thompson RL, Higgins CA, Howell JM (1991) Personal computing towards a conceptual model of utilization. *MIS Quarterly*.
- [15] Taylor S, Todd P (1995) Decomposition and crossover effects in the theory of planned behavior: a study of consumer adoption intentions. *International Journal of Research in Marketing*.
- [16] Tan M, Teo TSH (2000) Factors influencing the adoption of Internet banking. *Journal of Association of Information Systems*.
- [17] Ndubisi NO, Supinah R, Guriting P (2004) The extended technology acceptance model and internet banking usage intention. *International Logistics Congress Proceeding*, Turkey.
- [18] O'Cass A, Fenech T (2003) Web retailing adoption: exploring the nature of internet users Web retailing behavior. *Journal of Retailing and Consumer Services*.
- [19] Venkatesh V, Morris MG (2000) Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly*.
- [20] Adams DA, Nelson RR, Todd PA (1992) Perceived usefulness, ease of use, and usage of information technology: a replication. *MIS Quarterly*.
- [21] Doll WJ, Hendrickson A, Deng X (1998) Using Davis's Perceived Usefulness and Ease-of-use Instruments for decision making: a confirmatory and multi group invariance analysis. *Decision Sciences*.
- [22] Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*.
- [23] Copeland MT (1923) Relation of consumers' buying habits to marketing methods. *Harvard Business Review*.
- [24] Berry LL, Seiders K, Grewal D (2002) Understanding service convenience. *Journal of Marketing*.
- [25] Easterbrook G (1995) A moment on the earth: the coming of age of environmental optimism. Viking Press, New York, NY.
- [26] Lohse G, Spiller P (1998) Electronic shopping: how do customer interfaces produce sales on the Internet? *Commun*, pp: 81-87.
- [27] Degeratu AM, Rangaswamy A, Wu JN (2000) Consumer choice behavior in online and traditional supermarkets: the effects of brand name, price, and other search attributes. *International Journal of Research in Marketing* 17: 55-78.
- [28] Morganosky MA, Cude BJ (2000) Consumer response to online grocery shopping. *International Journal of Retail and Distribution Management* 28: 17-26.
- [29] Tanskanen K, Yrjola H, Holmstro J (2002) The way to profitable internet grocery retailing-six lessons learned. *International Journal of Retail and Distribution Management* 30: 169-178.
- [30] Colwell SR, Aung M, Kanetkar V, Holden AL (2008) Toward a measure of service convenience: multiple-item scale development and empirical test. *Journal of Services Marketing* 22: 160-169.
- [31] Moeller S, Fassnacht M, Ettinger A (2009) Retaining customers with shopping convenience. *Journal of Relationship Marketing* 8: 313-329.
- [32] Beauchamp MB, Ponder N (2010) Perceptions of retail convenience for in-store and online shoppers. *The Marketing Management Journal* 20: 49-65.
- [33] Yale L, Venkatesh A (1986) Toward the construct of convenience in consumer research. In Lutz RJ (Ed.), *Advances in Consumer Research*. Association for Consumer Research, Provo, UT, pp: 403-408.
- [34] Brown LG (1990) Convenience in services marketing. *The Journal of Services Marketing* 4: 53-59.
- [35] Seiders K, Voss GB, Godfrey AL, Grewal D (2007) SERVCON: development and validation of a multidimensional service convenience scale. *Journal of the Academy Marketing Science* 35: 144-156.
- [36] Srinivasan S, Anderson R, Ponnayolu K (2002) Customer loyalty in e-commerce: an exploration of its antecedents and consequences. *Journal of Retailing* 78: 41-50.
- [37] Gehrt KC, Yale LJ (1993) The dimensionality of the convenience phenomenon: a qualitative reexamination. *Journal of Business and Psychology* 18: 163-180.
- [38] Compeau DR, Higgins CA, Huff S (1999) Social cognitive theory and individual reactions to computing technology: a longitudinal study. *MIS Quarterly* 23: 145-58.
- [39] Hong W, Thong JYL, Wong WM, Tam KY (2001) Determinants of user acceptance of digital libraries: an empirical examination of individual differences and system characteristics. *Journal of Management Information Systems* 18: 97-124.
- [40] Sathye M (1999) Adoption of internet banking by Australian consumers: an empirical investigation. *International Journal of Bank Marketing* 17: 324-334.
- [41] Al-Somali SA, Gholami R, Clegg B (2009) An Investigation into the acceptance of online banking in Saudi Arabia. *Technovation* 29: 130-141.
- [42] Wang YS, Wang YM, Lin HH, Tang TI (2003) Determinants of user acceptance of Internet banking: an empirical study. *International Journal of Service Industry Management* 14: 501-519.
- [43] Gefen D, Karahanna E, Straub DW (2003) Trust and TAM in online shopping: an integrated mode. *MIS Quarterly* 7: 51-90.
- [44] Venkatesh V, Davis FD (2000) A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science* 46: 186-205.
- [45] Guriting P, Ndubisi NO (2006) Borneo online banking evaluating customer perceptions and behavioural intention. *Management Research News* 29: 6-15.
- [46] Jarvenpaa SL, Tractinsky N, Vitale M (2000) Consumer trust in an internet store. *Information Technology and Management* 1: 45-71.
- [47] Zikmund WG (2003) Business research methods, Cincinnati, Ohio: Thom-son/South-Western.
- [48] Coldwell D, Herbst F (2004) Business research. Cape Town, South Africa.
- [49] Frazer L, Lawley M (2000) Questionnaire design and administration: a practical guide. Brisbane: John Wiley and Sons, Australia.
- [50] Cooper DR, Schindler PS (2001) *Business Research Methods*. Irwin/McGraw-Hill, Singapore.
- [51] Nunnally J (1978) *Psychometric theory*. McGraw-Hill, New York.
- [52] Kristensen K, Eskildsen J (2010) Design of PLS-Based satisfaction studies. In: Esposito V, Chin WW, Henseler J, Wang H (eds.) *Handbook of partial least squares*, Springer handbooks of computational statistics, Heidelberg, pp: 247-277.
- [53] Agarwal R, Prasad J (1997) The role of innovation characteristics and perceived voluntaries in the acceptance of information technologies. *Decision Sciences* 28: 557-581.
- [54] Ajzen I, Fishbein M (1975) *Belief, Attitude, Intentions and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Boston.
- [55] Reimers V, Clulow V (2009) Retail centers: it's time to make them convenient. *International Journal of Retail and Distribution Management* 37: 541-562.
- [56] Rogers EM (1983) *Diffusion of innovations*. The Free Press, New York, NY, USA.

Feasibility Analysis of Directional-Location Aided Routing Protocol for Vehicular Ad-hoc Networks

Kamlesh Kumar Rana
Computer Science & Engineering
IIT (ISM) Dhanbad
Jharkhand, India
ranakamles@rediffmail.com

Sachin Tripathi
Computer Science & Engineering
IIT (ISM) Dhanbad
Jharkhand, India
var1285@yahoo.com

Ram Shringar Raw
Computer Science & Engineering
IGNTU Amarkantak
Madhya Pradesh, India
rsrao08@yahoo.in

Abstract—Vehicular Ad-hoc Network (VANET) is a multi-hop wireless ad-hoc network created by using mobile vehicles to transmit safety message for vehicle drivers. Since vehicles are mobile so they change their location frequently, therefore; robust data delivery is a challenging task in the VANET. Due to frequently network topology change characteristic, selection of a routing protocol in VANET is challenging task. In this paper performance of location-based routing protocols Directional-Location Aided Routing (*D-LAR*), Location-Aided Routing (*LAR*) and DIrectional Routing (*DIR*) are analyzed to decide best routing protocol for VANET. *LAR* protocol limits the route discovery area in the forward direction using GPS technology and *DIR* protocol uses direction information from the baseline drawn from the source and destination node. The *D-LAR* protocol uses concepts of the both *LAR* and *DIR* protocols. Using greedy forwarding approach *D-LAR* protocol selects next hop forwarding node in the forward direction of the communication range. Feasibility of *D-LAR* protocol has justified through simulation in NS2 using routing metrics such as node distribution at the border area of the communication range R , expected one hop distance $E(N(n,r))$, expected hop counts $E(H)$ between source and destination node, expected delay $E(\text{delay})$, routing overhead and packet loss. Through simulation work, it has shown *D-LAR* protocol performs better as compared to *LAR* and *DIR* protocol.

Keywords-VANET; DSDLAR; DLAR; LAR; SD

I. INTRODUCTION

Vehicular Ad-hoc Network (VANET) is a self-organized and decentralized wireless ad-hoc network uses mobile vehicles to transmit data packets throughout the network. VANET can be widely applied in so many fields as emergency deployments and community networking. In VANET, power and adequate storage are not an issue because VANET use vehicles as a node instead of other devices and they have sufficient energy and power for data processing and storage. One of the most important characteristics of VANET is dynamic topology [1] where vehicles move at very high speed on the road due to this they can change their network

topology changes rapidly. Due to highly dynamic nature, mobility model and location prediction play a very important role in designing of data dissemination in the network.

In VANET all nodes are mobile therefore an efficient routing is a fundamental and challenging task. The usage of physical location information of the nodes improves the efficiency of routing techniques of VANET. Location information of the nodes in VANET mainly leads to reduce routing overhead and increases packet delivery rate [2].

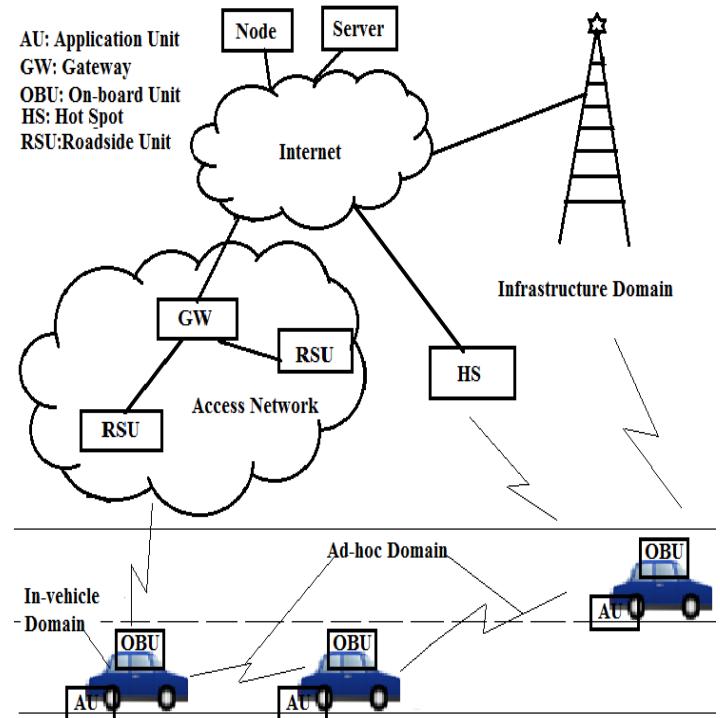


Figure1. VANET communication architecture

In VANET, routing protocols are required to transmit the data packet from a source node to destination node via a

number of intermediate nodes. The main purpose of routing protocol is to search a better route for data packet delivery to the correct destination in the network. For routing in *VANET* two types of routing protocols such as topology-based and location-based are used. Topology-based routing protocols contain information of the entire network in a routing table and location-based routing protocols contain location information of the neighbor nodes [3].

The major studies and research work done in *VANET* are mainly focused on traditional ad-hoc topology-based routing and location-based ad-hoc routing. Selection of routing protocols depend upon kind of network topology, therefore, it needs to study various routing protocols to select a suitable routing protocol for different kinds of the network topology in *VANET* [6]. There are some challenges and problems for the researcher to design a routing protocol for *VANET*.

1. How reliability of a routing protocol can be improved and along with the reduction of delay and retransmission of control overheads.
2. Scalability is another important issue in routing protocol due to varying network size in *VANET*; networks may be sparse or dense. In sparse network data packet must be carried by a node until the next-hop node is found in the dense network.
3. The behavior of the driver should also be considered to design the delay-bounded routing protocols since the carry-and-forward method is the main approach used for delivering the packets.
4. During designing of the routing protocol for a big city, interference caused due to the tall buildings present along the roadside should also be considered.
5. Security is one of the major issues; we need to further investigation and analyze the cooperation between inter-vehicular networks. With increasing number of vehicles on the road, the trust between these vehicles should also be sustained in order to have the smooth communication.

Day-by-day improvements in consumer's interest are ever increasing and it is an important research topic [7]. Vehicle-to-vehicle (V2V) and vehicle-to-roadside (V2R) and roadside unit-to-roadside unit (R2R) communications have become more popular in *VANET*. Most of the *VANET* research focused on urban and suburban roadways for dense network due to small inter-vehicle distance and terrain is not an important factor the fixed communication infrastructure is available [8].

Remaining parts of this paper is structured as follows: Literature survey and backgrounds on proposed DSDLAR protocol is discussed in section II. Section III presented proposed protocol in details. Section IV provides comparative

simulated results of the DSDLAR with DLAR and LAR protocol. Finally, the conclusion of the paper is given in section V.

II. LITERATURE SURVEY AND BACKGROUNDS

In *VANET*, nodes are mobile vehicles they move with varying speed on the road. Menouar et al. [12] proposed a location-based routing protocol for *VANET* named Movement Prediction Based Routing Concept (MOPRC). This protocol predicts current location of moving vehicles on the road and mobility model depend upon the nodes lifetime in a particular place useful to determine the route. This mobility model used current position, lifetime, direction and speed of the nodes for routing [12].

In *VANETs*, all nodes are highly movable so *VANETs* requires efficient routing protocols to find out the best path in the network for better performance of the network. Due to dynamic nature of nodes, it is very difficult to deliver data items from source to destination so an efficient routing protocol can perform well in all scenarios. For routing in *VANETs* vehicle speed, position, and network density are challenging issues. Vehicular ad-hoc networks for a highway depends upon speed and direction of vehicles thus *VANETs* requires customized routing protocols for better performance. Authors Kaleem, Hussain et. al. in [13] presented direction and relative speed based routing protocol for highways using a single hop packet forwarding approach. It selects the next hop using DARS of the vehicle.

Position based routing protocols are more suitable in *VANETs* as compared topology-based routing protocols due to advancement and usability of GPS device. Due to limitations of GPS, systems to collect current position of the nodes depend upon the beacon interval. So there a delay occurs during collecting the current position of nodes that forces routing protocol to use inaccurate position information of nodes that lead to low throughput and high overhead. Authors Siddharth Shelly and A. V. Babu in [14] proposed a Link Reliability Based Greedy Perimeter Stateless Routing for Vehicular Ad Hoc Networks that predicts the location of neighbor nodes of the sending node using speed and direction information provided in beacon packets during the beacon interval.

In *VANETs*, nodes are highly mobile on the road that causes network topology frequently changes and it decreases throughput and efficiency of the routing protocol. To improve the throughput and efficiency of routing protocol position based next-hop forwarding method has recommended for the linear and nonlinear network [15, 16]. The position based routing protocol is a useful protocol in multi-hop vehicular ad-hoc networks, due to the high mobility of nodes. Selection of next-hop node is an important factor to improve routing performance in the networks. Rao and Lobiyal in [17]

proposed a protocol that selects next-hop forwarding node based on the distance and link quality between the source and next-hop node. The expected delay and throughput also estimated for the proposed method.

In order to guarantee reliable data transmission and route calculation in the context of the proactive routing protocol in [18] proposed a Link State Aware Geographic Opportunistic Routing Protocol for *VANET* for multipoint relays selection. They considerably optimized end-to-end delay from sender to receiver nodes based on an updated estimation model of link lifetime correlated with a connectivity ratio by this proposal. Indeed this new approach could be the subject of many applications, where the delay of packets delivery is critical, namely, in the aerospace domain.

In sparse *VANET* number of vehicles be less so route maintenance is still more complex. T. Sivakumar in [19] proposed an efficient hybrid routing protocol for sparse *VANETs*. This is an on-demand routing protocol with proactive route maintenance (*OPRM*) using *RSUs* that repair the route in a sparse *VANET* by using roadside units (*RSUs*) in place of vehicles.

H. Takagi et. al. [20] has designed a reliable routing protocol for *VANET* using *GPSR* protocol, exploiting information about link reliability during the selection of one-hop forwarding vehicles. In this routing scheme node closer to the destination node and satisfies link reliability criterion will be chosen as next forwarder node. In addition, they have given an idea for probabilistic analysis of communication link reliability for one-dimensional *VANET* and this model used for the evaluation of the modified routing scheme. The routing method discussed in this ensures that most reliable nodes chosen for forwarding data packets and building a route from source to destination.

VANET is rapidly topology change wireless ad-hoc networks play a decisive role in public safety communication and commercial application. In *VANET* nodes are highly mobile due to this network topology rapidly changes accordingly routing of data is a challenging task. Position based routing protocols are becoming popular due to advancement and availability of *GPS* devices. One of the critical issues of *VANET* is frequent path disruptions caused by the high-speed mobility of vehicle that leads to broken links, which result in low throughput and high overhead. Authors in [21] presented the use of current location information of vehicles movements such as location, direction, and speed of vehicles to predict a possible link-breakage event prior to its occurrence.

III. DESCRIPTION OF PROPOSED PROTOCOL

Directional-Location Aided Routing [12] is an improvement of Location-Aided Routing [13] protocol based

on current location information of the nodes. To improve performance D-LAR protocol uses advantages of both routing protocols LAR and DIR. In D-LAR routing protocol efficiency of data delivery depends upon the current location information of the nodes obtained from the global positioning system or other location information services. After a specific time interval, each node finds own location information.

VANET is highly dynamic so to obtain correct current location information of nodes time interval should be small that turns into high communication overhead. High communication overhead in the network affects the overall network performance. Therefore, in a highly dynamic network a method is needed to find current location information and direction of nodes towards destination node with low communication overhead [12].

D-LAR protocol draws a baseline *SD* from source node *S* to destination node *D*, to select next hop forwarder node *D-LAR* routing protocol checks direction of each node from this baseline *SD*. The node closest to the baseline *SD* and moving in direction of the destination node *D* is selected as next hop forwarder node because it will give stable route in the direction of the destination node. Therefore, node selected as next hop forwarder node using this concept will reduce the average number of hop counts between source and destination node and data packet forwarding delay that increase network performance. Following figure 2 shows request zone and expected zone in the *D-LAR* protocol.

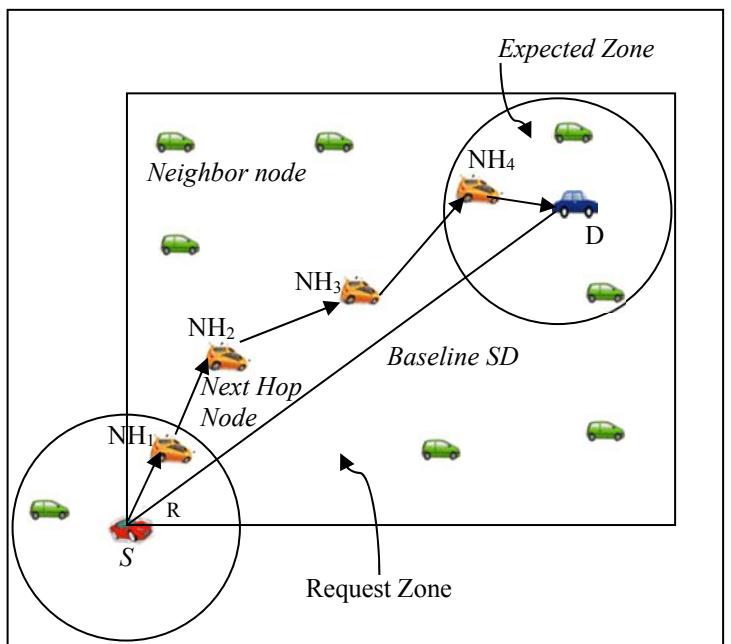


Figure 2. *D-LAR* routing scheme

Suppose, in fig. 2 current coordinate value of source node *S* and next hop node *NH₁* is (X_S, Y_S) and (X_{NH_1}, Y_{NH_1})

respectively. When next hop node NH_1 receives a route request $RREQ$ message from the source node S , it calculates two values first one distance d between source node S and next hop node NH_1 and second one angle θ from baseline SD as follows:

$$d = \sqrt{(S_x - NH_x)^2 + (S_y - NH_y)^2} \quad (1)$$

$$\theta = \tan^{-1} \left(\frac{S_y - NH_y}{S_x - NH_x} \right) \quad (2)$$

As shown in fig. 2, S and D represents source and destination node respectively. NH_1 is representing next hop node selected by the source node because it has the maximum distance from source node with the minimum angle from the baseline SD . In same fashion NH_2, NH_3 will be chosen as next hop node for further transmission of data packets in the network. Finally, next hop node NH_4 will deliver the packets to the destination node D . Compared to LAR and D-LAR protocol, the D-LAR protocol is more useful for a dense network environment such as city traffic scenario, where a plenty number of vehicles can make connectivity between vehicles.

A. Problem Statements

VANET comprises a large number of mobile vehicles moving randomly moving on the road. Each vehicle in *VANET* works as a transmitter because they participate in data transmission in the network. In *VANET*, each vehicle has a unique ID and fixed transmission range R . To select next hop forwarding node D-LAR uses greedy forwarding approach and selects a next hop forwarding node among the nodes at the border area of the communication range.

To achieve high network performance selection of next hop node must be appropriate because best next hop forwarding node reduces the average number of hop counts between the source and destination node and delay. This causes overall performance of the network increases. To select best next hop forwarding node *D-LAR* protocol restrict search area using the concept of expected zone and request zone. *D-LAR* is a location-based routing protocol useful in finding current location information and direction of nodes approaching destination.

Some routing metrics such as node distribution at the border area of the communication range R , expected one hop distance, expected hop counts between the source and destination node, expected delay, link lifetime and path duration are used to justify the performance of *D-LAR* protocol. Results obtained through NS2 simulation in result analysis section shown *D-LAR* protocol performs better as compared to *LAR* and *DIR* protocol.

B. Working Procedure of *D-LAR* Protocol

The working mechanism of the *D-LAR* protocol is as follows:

1. When source node S wants to communicate with a node in the network then source node recognizes request zone (RZ) and expected zone (EZ) as given in fig. 2.
2. RZ is a rectangular area which size is decided by source node S , RZ incorporates source node S and EZ diagonally opposite corner as shown in fig. 2.
3. Suppose, at time T_0 location of the source node S and destination node D is (X_S, Y_S) and (X_D, Y_D) and after time T_1 source node wants to communicate with the destination node D and knows its velocity V_D . Using details of the destination node, source node defines circular area EZ of radius $R = V_D(T_1 - T_0)$ centered at the location (X_D, Y_D) .
4. Source node S finds the distance of destination node D from its location using equation 3 and draws a baseline SD from itself to the destination node D .

$$SD = \sqrt{(X_S - X_D)^2 + (Y_S - Y_D)^2} \quad (3)$$

5. The baseline SD may either larger or smaller to the communication range R .
6. If $SD > R$, then source node S be outside of the expected zone as shown in fig. 3.

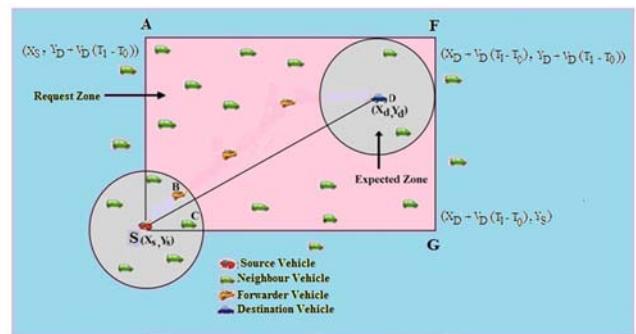


Figure 3. Source node outside of the expected zone

7. If $SD < R$, then source node be inside of the expected zone as shown in fig. 4.

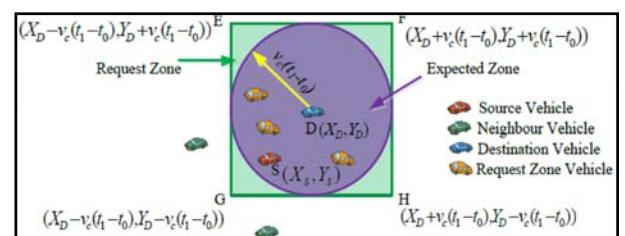


Figure 4. Source node inside the expected zone

8. Source node S initiates route discovery process by flooding route request $RREQ$ message to its all neighbor nodes.
9. Source node calculates distance and angle of each neighbor nodes using following equation 4.

$$D_I = \sqrt{(X_{N_I} - X_S)^2 + (Y_{N_I} - Y_S)^2} \quad (4)$$

$$\theta_I = \tan^{-1} \left(\frac{S_y - (NH_Y)_I}{S_x - (NH_X)_I} \right) \quad (5)$$

10. Source node S select a node as next hop forwarding node with maximum D_I and minimum θ_I from baseline SD .

C. Mathematical Analysis of D-LAR Protocol

In VANET, nodes are allowed to move within a specified area in the network at any speed, nodes are highly mobile that causes the link between nodes may be broken frequently. Therefore, to transmit data packets to the intended destination node an alternative path must be re-established immediately. In VANET, at any time either a new node can enter in the network or a node in the network can leave the network. The newly joined node in the network can establish the connection with the other nodes in the network and when a node leaves network then the connection established between nodes in the network will be broken. Due to this establishing connection between nodes and finding the average number of hop counts between the source and destination node is a challenging task. The number of nodes in the network directly affects the network performance because the higher number of nodes in the network excels probability of finding a best next hop node to forward the data packet to the intended destination node.

VANET is a highly dynamic network that causes links between nodes frequent breaks and in this situation existing path cannot deliver data packets unless an alternative path is not found by selecting a new node to forward the data packets immediately. Therefore, to verify the feasibility of *D-LAR* protocol mathematical analysis has done in the next subsection.

1. Probability of Node Distribution at Border Area of the Communication Range

The probability of node distribution at the border area of the communication range specifies the availability of nodes from that next-hop node can be selected for further data transmission. In fig. 5, it can be seen source node S would like to communicate with the destination node D in the network. The destination node D is out of the range of communication range of the source node S . Now, in this situation to complete the communication, intermediate next-hop are required. In this model, vehicle-to-vehicle (V2V) communication is considered that has no infrastructure or roadside unit (RSU) along the roadside.

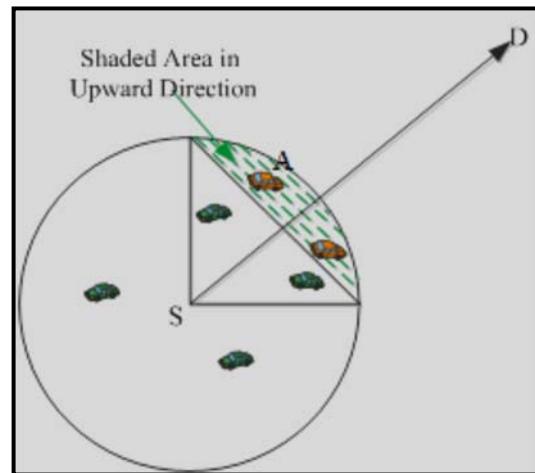


Figure 5: Distribution of nodes at the border area of the communication range.

Sensors are used on the vehicle to receive and transmit valuable traffic-related information of the network for the vehicle driver so that driver can take appropriate decision for driving the vehicle. Here transmission range of each vehicle is assumed to equal denoted by R and communication link between two vehicles depend only on the distance between them. In VANET, each vehicle is able to obtain own current location and velocity information with the help of GPS and other built-in digital roadmaps.

Suppose nodes in the network follow Poisson distribution process and K nodes arrive in the shaded area shown as in figure 3.2. The arriving nodes in the shaded area can be calculated as follows:

$$P(K) = \sum_{N=K}^{\infty} \binom{N}{K} (p)^K (1-p)^{N-K} \left(\frac{(0.352 * \rho * X^2)^N * e^{-0.352 * \rho * X^2}}{N!} \right) \quad (6)$$

Therefore,

$$P(K) = \frac{(0.352 * \rho * p * X^2)^N * e^{-0.352 * \rho * p * X^2}}{N!} \quad (7)$$

Now, the probability of at least K neighbor nodes consequently in the given area as shown in the above figure 4.2 can be found as follows:

$$P(K_a) = 1 - \sum_{i=0}^{K-1} \frac{(0.352 * \rho * p * X^2)^i * e^{-0.352 * \rho * p * X^2}}{i!} \quad (8)$$

Now, we are able to calculate selection probability of at least one node in the shaded area within the transmission range R with the help of equation 3 as follows:

$$P = 1 - P(X = 0) \quad (9)$$

$$P = 1 - e^{-0.352 * \rho * p * R^2} \quad (10)$$

2. One Hop Expected Distance

As shown in the fig. 5, node S represents source node and node A represents border node at the border area of the communication range. The border node A can be used as a next-hop forwarding node positioned at the maximum distance within transmission range R . Suppose there are N neighboring nodes of the source node, S in the forward area towards the destination node, D .

Let $N-1$ nodes out of N nodes are within shaded area and N^{th} node is at maximum distance or closer to border of the sender's transmission range. Suppose, $d(S, N_l)$ denotes the distance between source and l^{th} node at border area of the communication range and Cumulative Distribution Function (CDF) of $d(S, N_l)$ is $F(r)$ as following:

$$\begin{aligned} F(x) &= p\{d(S, N_l) \leq x\}; x \in (-\infty, \infty) \\ F(x) &= P[d_1 \leq x, d_2 \leq x, d_3 \leq x, \dots, d_n \leq x] \\ F(x) &= \prod_{i=1}^n P[d_i \leq x] = \left(\frac{x}{R}\right)^n \end{aligned} \quad (11)$$

Suppose there are N nodes at the border area of the communication range, then there may be $n \cdot (n-1)/2$ links between the sender node and them no longer than communication range R as $N(n, R)$. Where $N(n, R)$ is a random variable represents the distance of border nodes whose expected value can be expressed as:

$$E(N(n, r)) = \frac{n \cdot (n-1)}{2} \cdot \int_r^R F(x) dx \quad (12)$$

$$E(N(n, r)) = \frac{n \cdot (n-1)}{2} \int_r^R \left(\frac{x}{R}\right)^n dx \quad (13)$$

$$E(N(n, r)) = \frac{n \cdot (n-1)}{2 \cdot R^n} \left[\frac{x^{n+1}}{n+1} \right]_r^R \quad (14)$$

$$E(N(n, r)) = \frac{n \cdot (n-1)}{2 \cdot R^n} \cdot \left[\frac{R^{n+1} - r^{n+1}}{n+1} \right] \quad (15)$$

3. Expected Hop Counts between Source and Destination Node

In the network scenario node distribution follows Poisson process, neighbour distance distribution function for a point lies in the Euclidean distance space, R^d can be defined as:

$$F(y) = 1 - P(N(b(o, y)) = 0 | o) \quad (16)$$

$$F(d) = 1 - (e^{-|N(b(o, y))|}) \quad (17)$$

Where $P(N(b(o, y)) = 1 | o)$ is conditional probability that shows there is at least one point out of N points located in

the area $b(o, x)$. Area bounded by the $b(o, x)$ can be defined as $b(o, y) = \rho R^2$. Substituting value of $(N(b(o, y))$ in the equation 1, we will get as:

$$F(y) = 1 - e^{-\rho \pi R^2} \quad (18)$$

Similarly,

$$f(y) = \frac{d}{dy} F(y) \quad (19)$$

$$f(y) = \frac{d}{dy} (1 - e^{-\rho \pi R^2}) \quad (20)$$

$$f(y) = (2\rho \pi R \cdot e^{-\rho \pi R^2}) \quad (21)$$

Thus, the probability of one hop count can be calculated as:

$$P1 = 2\rho \pi \int_0^R R \cdot e^{-\rho \pi R^2} dR \quad (22)$$

Here $\rho \pi x^2$ is the area of the circle and in our proposed model, we have considered the only quarter area of the circle, so we can replace $\rho \pi R^2$ with $\frac{\rho \pi R^2}{4}$.

Now,

$$P1 = 2\rho \pi \int_0^R e^{-\frac{\rho \pi R^2}{4}} dR \quad (23)$$

Put $R^2 = y$; $2R \cdot dR = dy \Rightarrow R \cdot dr = \frac{1}{2} dy$; Now limit will be as $x=0, y=0, x=R, y=R^2$. So, $P1$ can be written as following:

$$P1 = 2\rho \pi \int_0^y e^{-\frac{\rho \pi y}{4}} \frac{1}{2} dy \quad (24)$$

$$P1 = \rho \pi \int_0^{R^2} e^{-\frac{\rho \pi y}{4}} dy \Rightarrow P1 = 4 \cdot \left[1 - e^{-\frac{\rho \pi R^2}{4}} \right] \quad (25)$$

Here, $\frac{\rho \pi R^2}{4}$ represents the total number of nodes at the border area of the communication range so we can write $P1$ as follows:

$$P1 = 4 \cdot [1 - e^{-N}] \quad (26)$$

Therefore, the probability of two-hop count can be obtained as follows:

$$P2 = \rho \pi \int_{R^2}^{4R^2} e^{-\frac{\rho \pi y}{4}} dy \quad (27)$$

$$P2 = \int_{R^2}^{4R^2} 4 \cdot \left[-e^{-\frac{\rho \pi y}{4}} \right]_{R^2}^{4R^2} dy \quad (28)$$

$$P2 = 4 \cdot [e^{-N} - e^{-4N}] \cdot [1 - e^{-N}] \quad (29)$$

For probability of the consequent hop counts the above equation can be generalized as follows:

$$P_t = 4 \cdot [e^{-N(t-1)^2} - e^{-Nt^2}] \cdot [1 - e^{-N}]^{t-1} \quad (30)$$

From the above equation, the average number of hop counts between the source and destination node can be calculated as follows:

$$E(H) = \sum_{H=1}^t H.P(t) = P_1 + 2P_2 + 3P_3 + 4P_4 + \dots + mP_m \quad (31)$$

$$E(H) = \sum_{H=1}^t 4H. [e^{-N(H-1)^2} - e^{-NH^2}]. [1 - e^{-N}]^{(H-1)} \quad (32)$$

4. Expected Delay

To improve the network performance in VANET, it is required to select a suitable next-forwarder hop with the suitable path in the network to forward data packets. To minimize the delay during data transmission, suitable routing protocol such as position based routing protocol communicates packets using radio waves as earliest. Since, in VANET, roads can be used as a medium for vehicular nodes through which the packet has to be transferred, therefore, the road with maximum velocity is selected first.

During data transmission, all the routing protocols in VANET assumes that smart vehicular nodes are furnished with the computing device, sensors, digital maps, and advanced information processing tools. Digital map in the vehicle provides street and lane level map for drivers and traffic-related information such as traffic density on the road, direction of nodes, position and velocity of vehicular nodes on the roads at disparate times of the day. Total delay is the time required to transmit data packets from source node to destination node. Therefore, the expected delay between two hops can be defined as:

$$T_{delay} = \text{Probability of at least one node * Speed} \\ + \text{Probability of Nonodes * Speed}$$

In the sender's transmission range, the probability of at least one node can be given as:

$$P(x=1) = (1 - e^{-\rho R}) \quad (33)$$

Similarly, the probability of no node in the transmission range is:

$$P(x=0) = e^{-\rho R} \quad (34)$$

Therefore, the expected delay can be written as follows:

$$T_{delay} = (1 - e^{-\rho X}) \cdot \frac{E(k)}{S} + e^{-\rho X} \cdot \frac{E(k)}{S} \quad (35)$$

Where,

$E(k)$ = Expected Distance between two hops

X = Communication Range of the Node

S = Speed of vehicles

ρ = Node density in the network

5. Expected Progress Distance

To find out shortest path between the source and the destination node, the expected maximum distance $E(X_{max})$ between source node and next-hop forwarding node can be used. Supposed, P_{Size} and W represent the size of the delivered packets on the network and link bandwidth respectively. The expected transmitted packets E_{TX} are used to maximize network performance in term of throughput and link quality in Wireless Ad-Hoc Network. For better performance of a wireless network the expected transmitted packets E_{TX} should be smaller for the link. Normally it is the sum of the E_{TX} value of each link along the path. But in the case of location-based Greedy forwarding method expected transmitted packets E_{TX} is measured by using periodically broadcast control message frequently sent on the network. Suppose p is the probability of successfully delivered packet and $q=1-p$ is the probability of unsuccessful to delivered packets. The successful expected transmitted packets by a node M to the next-hop forwarding node can be given as:

$$E_{TX} = \sum_{N=1}^{\infty} M.p^M(1-p)^{M-1} \quad (36)$$

$$E_{TX} = \left(\frac{1}{(1-p)}\right) \quad (37)$$

The expected transmission time (E_{TT}) of a link can be given as follows:

$$E_{TT} = E_{TX} \cdot \left(\frac{P_{Size}}{W}\right) \quad (38)$$

$$E_{TT} = \left(\frac{1}{(1-p)}\right) \cdot \left(\frac{P_{Size}}{W}\right) \quad (39)$$

To find out the expected progress distance of a next forwarding node we can relate both expected maximum distance $E(X_{MAX})$ of next forwarding node obtain in equation 7 and expected transmission time E_{TT} as following:

$$E_{PD} = \frac{E(X_{MAX})}{E_{TT}} \quad (40)$$

$$E_{PD} = \left(\frac{\frac{N}{R^N} \cdot \left(\frac{R^{N+1} - r^{N+1}}{N+1}\right)}{\left(\frac{1}{(1-p)}\right) \cdot \left(\frac{P_{Size}}{W}\right)}\right) \quad (41)$$

$$E_{PD} = \frac{N}{R^N} \cdot \left(\frac{R^{N+1} - r^{N+1}}{N+1}\right) \cdot \left(\frac{W \cdot (1-p)}{P_{Size}}\right) \quad (42)$$

6. Packet Loss Rate

The packet loss rate is a ratio between the total number of delivered data packets by the source node and number of data packets received by the destination node. Suppose, source node S sends M data packets to a CNH nodes, but it received only M_{recv} data packets. Then the packet loss rate of a link can be given as:

$$Pkt_{loss} = \left(\frac{P_{recv}}{P_{Delivered}} \right) 100\% \quad (43)$$

IV. SIMULATION AND RESULT ANALYSIS

VANET uses the mobile vehicle as nodes to transmit data packets in the network and they can move on the road with varying speed. The simulation model of the network with variable mobile nodes 10-100 placed randomly within a 800 x 800 m² area. There are many different ways to measure network performance and can be modeled using Network Simulator. Performance of D-LAR protocol is evaluated using network simulator NS2 in term of nodes probability, one hop node distance, average number of hop counts, routing overhead, packet loss rate, delay and throughput. The parameters used in simulation summarized in table 1. Sample topology of NS2 is shown in fig. 6.

Table1. Used simulation parameters

Parameters	Default Values	Parameters	Default Values
Simulation area	800m X 8000m	Traffic type	CBR
Communication range	200-300m	CBR interval	0.5 s
Number of nodes	10-100	Hello interval	1 s
Node speed	5-40 m/s	Window size	15 s
Packet size	1000 bytes	Simulation time	180 s

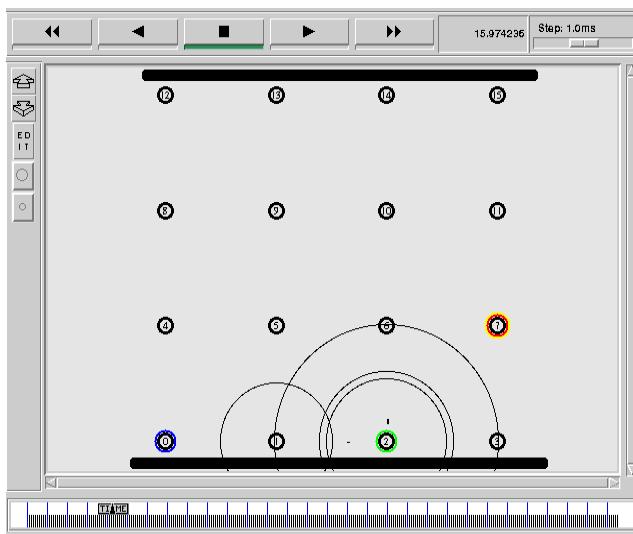


Figure 6. Sample of network topology using 16 nodes

In this simulation, any four nodes act as sender nodes with the traffic rate 0.5 seconds. As shown in fig. 3, node-0, node-1, node-2, and node-3 are the sender nodes. The

receiving nodes are node-12 and node-13. As shown in fig. 5, colored circle nodes indicate the transmission of packets with the large circle in animation topology. The red colored circle node-7 indicates congestion of traffic at a time of simulation.

A. Probability of Nodes Distribution

Node distribution specifies the number of nodes available at the border area of the communication range to select next hop node for further transmission of data packets if the destination node is out of reach of the sending node. The higher number of nodes in the network increases selection probability of the best next hop node. The best next hop node increases the overall performance of the network.

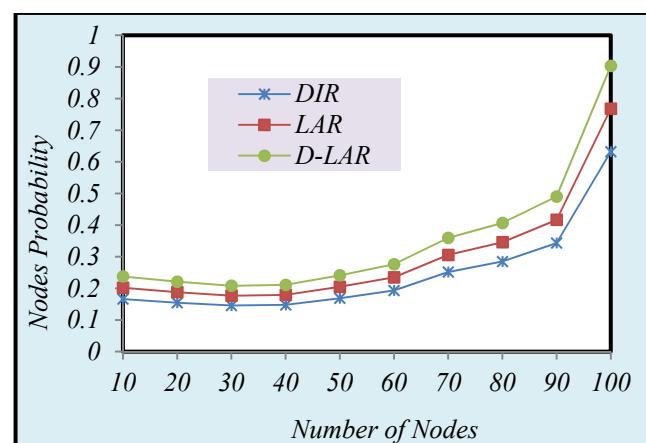


Figure 7(a). Probability of nodes vs. nodes

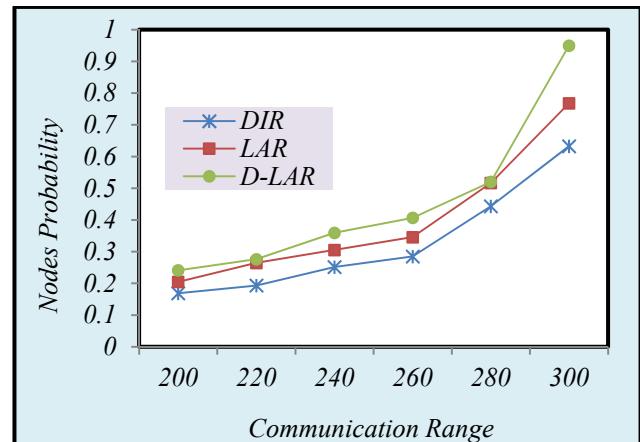


Figure 7(b). Probability of nodes vs. communication range

Fig. 7(a) and 7(b) depict the probability of nodes at the border area of the communication range vs. the number of nodes and communication range R of the nodes. It can be observed in both fig. 7(a) and fig. 7(b), the probability of nodes increases as the number of nodes in fig. 7(a) and communication range in fig. 7(b) increases. The higher probability of nodes increases selection probability of the best

next hop node, the best next hop node increases overall network performances. The probability of nodes at the border area of communication range is higher in the *D-LAR* protocol so it will perform better as compared to *DIR* and *LAR* protocol.

B. Routing Overhead

Routing overhead is a number of control packets required to discover the best route in the network for data transmission. It is an important parameter to measure the network performance; for better performance of a network control overhead should be low that causes node will consume less time to find the best route in the network.

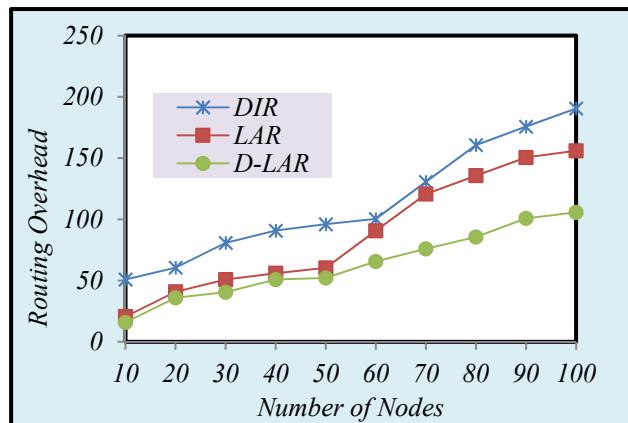


Figure 8(a). Routing overhead vs. nodes

Fig. 8(a) shows the routing overhead of *DIR*, *LAR* and *D-LAR* protocols vs. the number of nodes. It can be seen in figure routing overhead of each protocol increases as the number of nodes increases in the network. Reason behind this is that as the number of nodes increases the connection between sending and border nodes increases thus transferring of the routing packets between the sending and border nodes increases. Control packet overhead of routing protocol *D-LAR* is about 43.157% while in *LAR* and *DIR* is about 58.037% and 69.395% that are high as compared to *D-LAR* protocol.

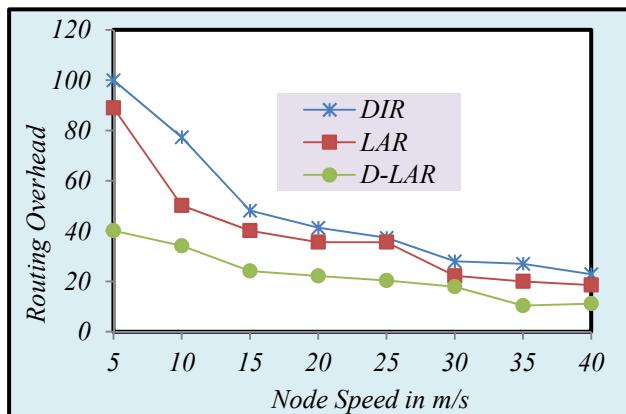


Figure 8(b). Routing overhead vs. node speed

Fig. 8(b) depicts the routing overhead vs. the nodes speed and it decreases as node speed increases. The reason behind this is that when the speed of nodes increase link duration between nodes decreases so they carry less number of control packets. Routing overhead of *D-LAR* is about 45.868% and in *LAR* and *DIR* is about 66.885% and 73.219% that are higher as compared to *D-LAR* protocol.

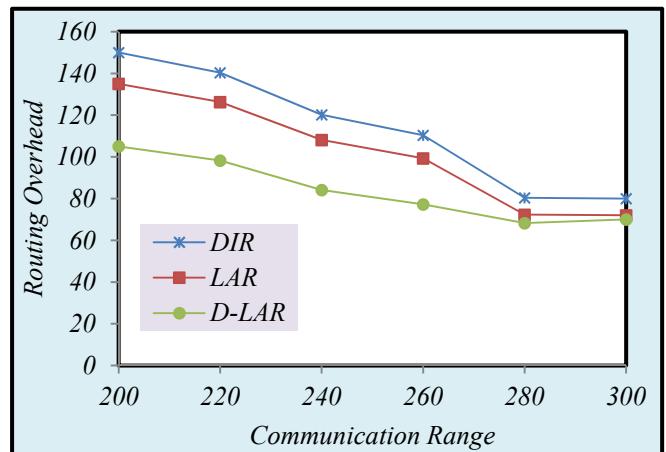


Figure 8(c). Routing overhead vs. communication Range

Fig. 8(c) shows routing overhead with respect to the communication range of nodes. It can be observed in the figure routing overhead decreases as node communication range increases. The reason behind this is that when communication range of a node increases, it requires more time for carrying and forwarding data packets from source to destination node, so the number of transmitted data packets decreases. Routing overhead of the *D-LAR* is about 75.868% and in *DIR* and *LAR* is about 84.885% and 92.219% that are higher as compared to *D-LAR* protocol.

C. Routing Protocol Packet Drop Rate

Packet drop occurs when data packets travel across the network fail to reach their destination. Usually, packet drop occurs due to poor connectivity between nodes and network congestion. Packet drop rate is an important parameter to measure network performance, for better network performance packet drop rate in the network should be low. Packet drop is measured as a percentage of packets lost with respect to packets sent across the network.

Fig. 9 (a) shows packet drop rates with varying speed of the nodes from 5 to 40 m/s. As shown in the figure, packet drop rate increases as the speed of the nodes increase reason behind this is that network stability decreases as nodes speed increases. Packet drop rate *D-LAR* protocol is 28.63% while in *LAR* and *DIR* is about 67.47% and 77.59%. Thus it can be said *D-LAR* will perform better.

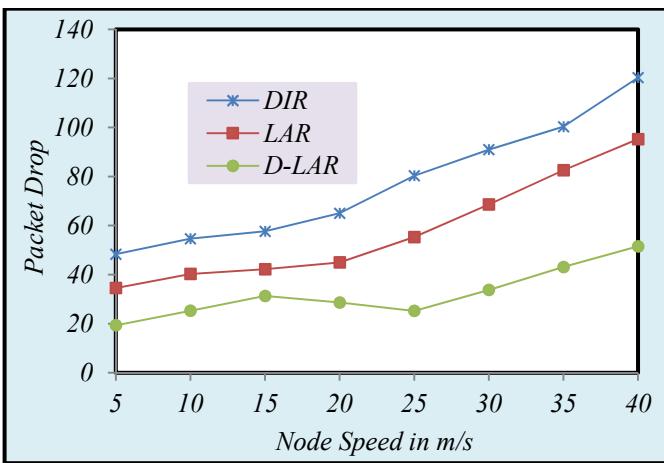


Figure 9(a). Packet drop rate vs. node speed

Fig. 9(b) depicts packet drop with varying number of nodes in the network from 200 to 300. As shown in figure packet drop rate decreases as number of nodes increases in the network because connection strength among the nodes increases as number of nodes increases in the network. Packet drop rate of routing protocol *D-LAR* is about 25.345% while in *LAR* and *DIR* is about 55.99% and 64.513% that are higher as compared to *D-LAR*.

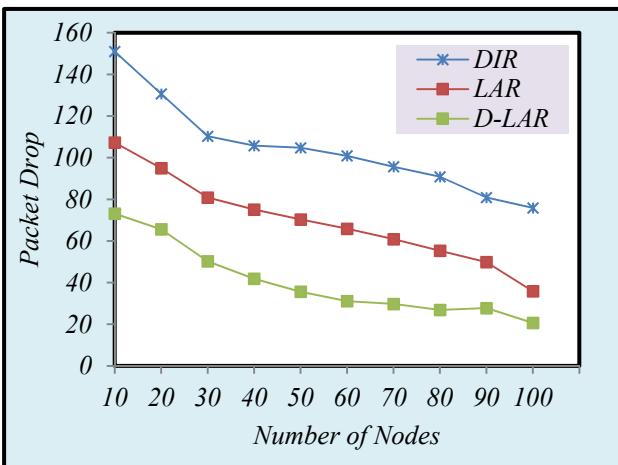


Figure 9(b). Packet drop rate vs node

D. Routing Protocol Delay

Delay is an amount of time required to transmit data packets from a node to another node in the network and it is an important parameter to measure the performance of the network. For better performance of the network and to reduce the large number of accidents delay should be low so that data packet can reach earliest at the destination.

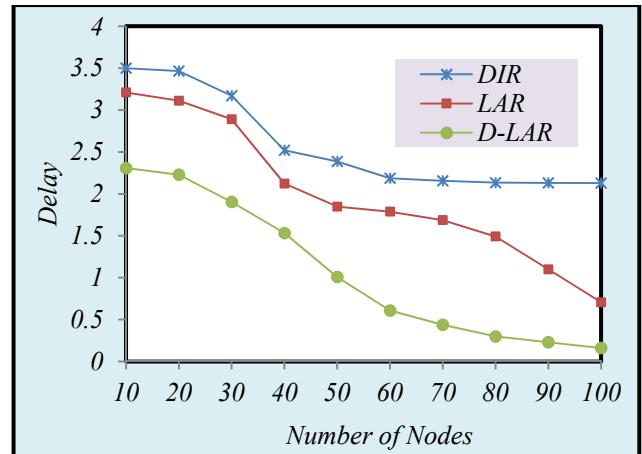


Figure 10(a). Delay vs. nodes

Fig. 10(a) shows routing protocol end-to-end delay with respect to varying nodes from 10 to 100. It can be observed in figure 10(a) routing protocol delay decreases as numbers of nodes increases. The reason behind this is that link quality among the nodes increases as nodes increases in the network. Delay in *D-LAR* routing protocol is nearly 21.31% while in other hand delays in routing protocol *LAR* and *DIR* is about 43.235% and 58.386% respectively.

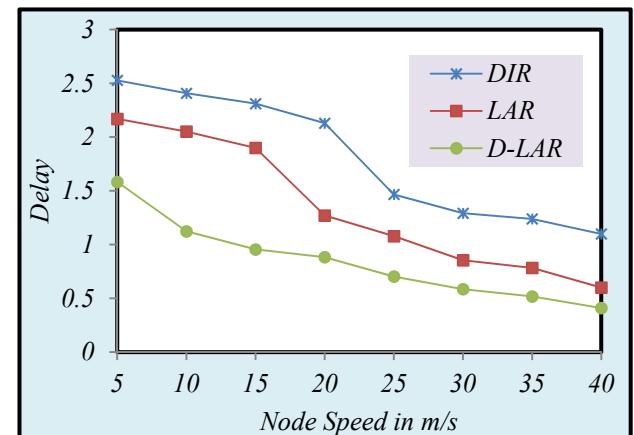


Figure 10(b). Delay vs. node speed (m/s)

Fig. 10(b) shows end-to-end delay that is an amount of time spent during successful data transmission from the source to the destination node. We can see in figure end-to-end delay declines as vehicle speed increases. The reason behind this is that the time required carrying and forwarding data packets decreases as vehicle speed increases. We can see the delay in *D-LAR* routing protocol is nearly 48.88% while in other hand delay in routing protocol *LAR* and *DIR* is about 76.51% and 83.82% respectively. For better performance of the network, the delay should be low thus we can say *D-LAR* will perform better.

E. One Hop Distance

One hop distance is a distance between the source node to next hop node and it is an important routing metric. For better performance of the network on hop distance should be high because the higher value of one hop distance reduces the average number of hop counts between source and destination node. The minimum number of average number of hop count between the source and destination node takes less time to deliver data packets.

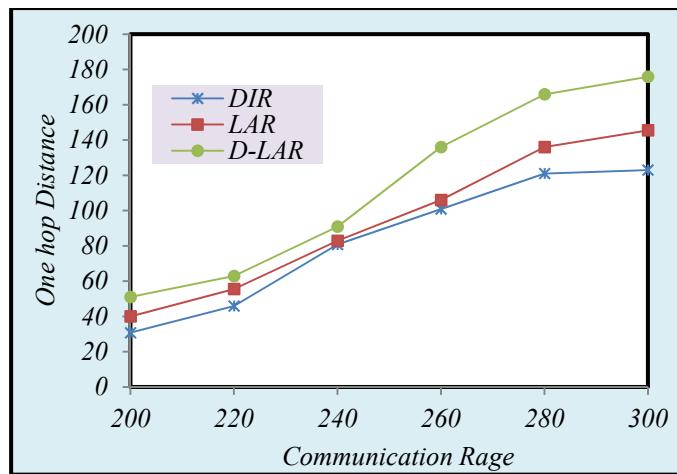


Figure 11(a). One hop distance vs. communication range

Fig. 11(a) depicts one hop distance versus communication range of nodes, it can be observed in figure one hop distance increases as communication range increases. One hop distance in the *D-LAR* protocol is higher as compared to *DIR* and *LAR* routing protocol that will perform better.

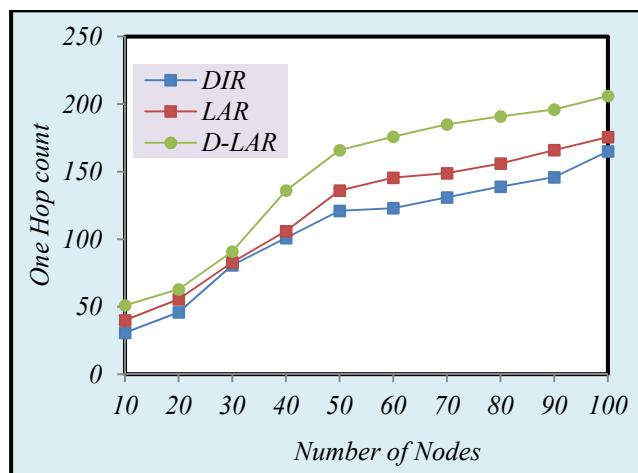


Figure 11(b). One hop distance vs. nodes

Figure 11(b) shows one hop distance with respect to varying nodes. One hop distance increases as the number of nodes increases in the network and it is higher in the *D-LAR* routing protocol as compared to *DIR* and *LAR* protocol.

F. Average Number of Hop counts

The average number of hop counts represents the number of intermediate nodes required to deliver data packets to the intended destination node. The higher number of intermediate nodes decreases network performance because it will take more time to deliver data packets at destination node.

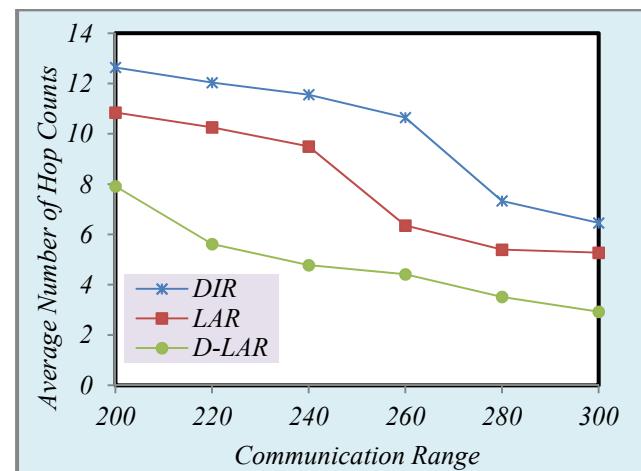


Figure 12. Average number of hop counts vs. communication range

Fig. 12 is the graphical representation of the average number of hop counts with respect to communication range. In the figure, it can be observed in figure average number of hop counts decreases as communication range increases. In *D-LAR* protocol value of the number of hop counts is lower as compared to *LAR* and *DIR* protocol.

V. CONCLUSION

Simulation results obtained through NS2 have shown the performances of the location-based routing protocols in different scenarios and with distinct performance metrics like nodes distribution, one hop distance, average number of hop counts, routing overhead, packet loss and delay. It has been found that *D-LAR* protocol performance is better in the group of location-based routing protocols.

REFERENCES

- [1] Kamlesh Kumar Rana, Sachin Tripathi and Ram Shringar Rao, "Analysis of Expected Progress Distance in Vehicular Ad-hoc Network using Greedy Forwarding" 11th INDIACom; INDIACom-2017; IEEE Conference ID: 403532017 4th International Conference on "Computing for Sustainable Global Development", 01 – 03 March 2017.

- [2] Ankita M. Shendurkar, Nitin R. Chopde, " A Review of Position Based Routing Protocol in Mobile Ad-Hoc Networks", International Journal of Advanced Research in Computer Engineering & Technology, Volume 3 Issue 6, June 2014.
- [3] Kamlesh Kumar Rana, Sachin Tripathi and Ram Shringar Rao, " VANET: Expected Delay Analysis for Location Aided Routing (LAR) Protocol", BIJIT - BVICAM's International Journal of Information Technology, Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi, India, 2016.
- [4] Rahem, Ismail, Ariffidris and Aymen, "A Comparative and Analysis Study of VANET Routing Protocols", Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, Vol. 66 No.3, 31st August 2014.
- [5] Kamlesh Kumar Rana, Sachin Tripathi, Ram Shringar Rao, " Analysis of Expected Hop Counts and Distance in VANETs", International Journal of Electronics, Electrical and Computational System IJEECS ISSN 2348-117X Volume 5, Issue 4 April 2016.
- [6] Hongyu Tu, Lei Peng, Huiyun Li, FalinLiu, "GSPR-MV: a Routing Protocol Based on Motion Vector for VANET", ICSP, 2014.
- [7] Shuai Yang, Rongxi He, Sen Li, Bin Lin, Ying Wang," An Improved Geographical Routing Protocol and its OPNET-based Simulation in VANETs", 7th International Conference on Bio Medical Engineering and Informatics,2014
- [8] Rupesh and Rao, "Directional Greedy Routing Protocol (DGRP) in Mobile Ad-hoc Networks", IEEE Computer Society, International Conference on Information Technology, 2008.
- [9] G. V. Rossi, K. K. Leung, and A. Gkelias, "Density-based optimal transmission for throughput enhancement in vehicular ad-hoc networks," Communications (ICC), 2015 IEEE International Conference on, pp. 6571-6576, 2015.
- [10] Karp, B. and Kung, H. T., "GSPR: greedy perimeter stateless routing for wireless networks", In Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, MobiCom 00, ACM, New York, NY, pp. 243-254, Agust-2000.
- [11] Bouamoud Bachir, Ouacha Ali et al., "Proactive Schema Based Link Lifetime Estimation and Connectivity Ratio", Hindawi Publishing Corporation The Scientific World Journal, Volume 2014, Article ID 172014, 2014.
- [12] Menouar et al., "Movement Prediction-Based Routing (MOPR) Concept for Position-Based Routing in Vehicular Networks ", Conference: Vehicular Technology Conference, 2007.
- [13] Kaleem, Hussain et al, "A direction and relative speed (DARS)-based routing protocol for VANETS in a highway", Taylor Francis, Journal of the Chinese Institute of Engineers, 2014.
- [14] Siddharth Shelly and A. V. Babu, "Link Reliability Based Greedy Perimeter Stateless Routing for Vehicular Ad Hoc Networks", International Journal of Vehicular Technology, Volume 2015, Article ID 921414, 2015.
- [15] G. V. Rossi, K. K. Leung, and A. Gkelias, "Density-based optimal transmission for throughput enhancement in vehicular ad-hoc networks," Communications (ICC), 2015 IEEE International Conference on, pp. 6571-6576, 2015.
- [16] K. Prasanth, K.Duraiswamy, K.Jayasudha, C.Chandrasekar, "Improved packet forwarding approach in vehicular ad-hoc networks using RDGR algorithm", International Journal of Next Generation Network, Vol.2, No.1, March 2010
- [17] R. S. Rao and D. K. Lobiyal, " Throughput and delay analysis of next-hop forwarding method for nonlinear vehicular ad-hoc networks", International Journal of Ad-Hoc Networking System, Vol.2, No.2, April 2012.
- [18] Cai et al., " LSGO: Link State aware Geographic Opportunistic routing protocol for VANETs", EURASIP Journal on Wireless Communications and Networking, Springer Open journal 2014.
- [19] T. Sivakumar , "OPRM: an efficient hybrid routing protocol for sparse VANETs", International Journal Computer Applications in Technology, Vol. 51, No. 2, 2015.
- [20] H. Takagi and L. Kleinrock, "Optimal Transmission range for randomly distributed packet radio terminals" IEEE Transactions on Communications 32 (3), pp. 246–257, 1984.
- [21]. Boukerche, Oliveira et. al., "Vehicular ad-hoc networks: A new challenge for localization-based systems", Computer Communication, Amsterdam: Elsevier, 2008.
- [22] Chi Trung Ngo, Hoon Oh, "A Link Quality Prediction Metric for Location based Routing Protocols under Shadowing and Fading Effects in Vehicular Ad Hoc Networks", Procedia Computer Science, Volume 34, Pages 565-570, 2014.

San Bernardino Symphony Orchestra and Exploring the Use of Mobile Applications by Symphony Orchestras

Abdullah Almusallam (abdull-musallam@hotmail.com), Evelia Avila, Prabhjeet Grewal and Abdulmajid Alnoamani (mj7557@hotmail.com).

Abstract

This report consists of an overview of the IT consultancy with the San Bernardino Symphony Orchestra (SBSO) and the implementation of the selected IT solution by the Client. The solution aimed to address the Client's need to attract more concert attendees and donors by providing a tool that leverages the power of data and geographic information systems (GIS) to identify untapped areas for potential outreach. In addition to this, the project team is interested in examining the use of mobile applications (apps) by symphony orchestras as another tool to enhance audience engagement and participation. We will evaluate the viability and effectiveness of mobile app usage for the purpose of increasing concert attendance. Data on industry IT expenditures, concert attendance, and mobile app usage will be analyzed to determine recommendations for symphony orchestras like SBSO.

Keywords: Information technology, GIS, non-profit organizations, symphony orchestras

1. Introduction

The art of classical music and symphony orchestra performances are pastimes enjoyed by many. There exists a diverse population of symphony orchestra performance attendees and patrons (Hager & Winkler, 2012; Ostrower, 2005). Regrettably, symphony orchestras have also experienced a decline in recent years and face persistent economic hardships. According to Rosen (2017), two major longitudinal studies on orchestras revealed a 13% decline in the proportion of the population that attended classical concerts between 2002 to 2008. More surprisingly, a 39% decline in attendance of college-educated adults was observed the same six-year period. Traditionally, education attainment has been known to be positively correlated with participation in the performing arts (Hager & Winkler, 2012).

This opens the gates to other areas of research that seek to understand factors that contribute to concert non-attendance for interested audiences (National Endowment for the Arts, 2012). In a separate study, Rosen (2017) examines the churn rate for orchestra performances and found that the “aggregated churn rates across orchestras in nine major markets was 80% [sic]” (p. 21). In other words, only two out of every ten first-time attendees would return for another performance. Among several reasons cited, the most popular reasons first-time attendees did not return were the aggressive fundraising attempts that ensued their visit, poor parking, and the inability to exchange tickets (Rosen, 2017).

Another key issue that lends itself to the economic difficulties faced by the performing arts is the very nature of being non-profit organizations. According to Baumol and Bowen (1965), the dilemma often experienced by non-profits is the ability to easily expend new money as soon as it becomes available while experiencing difficulty financing other projects. Thus, non-profits constantly find themselves on the brink of “financial catastrophe” (Baumol & Bowen, 1965, p. 497).

In recent years, the proliferation of technological advancements has afforded a wide array of opportunities across various industries. However, access to such technology is still a challenge for certain industries, including non-profit organizations with limited financial resources. Non-profit organizations, like symphony orchestras, have the potential to propel to greater heights by leveraging the use of technology and overcoming these financial challenges. There is a growing interest in the utilization of technology in the areas of performing arts education and audience engagement. Certain technologies have been found to perpetuate interest-driven arts learning (Peppler, 2013). Moreover, studies have also shown changes in audience engagement due to technology and arts delivery methods (Arts Index, 2016).

The focus of this report will center around symphony orchestras and IT. We will begin with an overview of the IT consultancy and solution implementation for the San Bernardino Symphony Orchestra (SBSO) and expand on existing research in the realm of mobile application usage and potential for enhanced audience attendance.

2. Overview and Client Needs

In the heart of the Inland Empire, the SBSO serves a diverse population of symphonic enthusiasts. The mission of the SBSO is to “foster a love of music, excite the spirit, and enrich [their] diverse community and region through live orchestral performances and music education” (San Bernardino Symphony Orchestra, 2017). The SBSO hosts five annual concerts at the historic California Theatre in downtown San Bernardino in addition to performing a variety of concerts throughout the region. Ticket sales, however, only represent approximately 20% of their annual income. In 2016, the total revenue for SBSO amounted to \$493,186, nearly a 22% decline from the year prior. Donations and cash contributions accounted for 71% of the total revenue (Internal Revenue Service, 2016). Evidently, the generous contributions from patrons are key in supporting the SBSO with its mission. Presently, the SBSO is seeking to increase their concert attendance and donors. Aside from supporting the musical legacy of the orchestra, donors also help perpetuate the various music education programs throughout the local schools. Therefore, acquiring new donors is vital for the continued support of the SBSO. The organization does not have any software or system in place which could help them locate specific areas to concentrate in order to enlarge their potential donors.

2.1 IT Project Solution

The goal for the project management team was to provide the SBSO with the optimal solution which would help them identify untapped areas with potential to attract new concert-goers and help them elevate their donor base. Research for this project was conducted in various areas of IT as well as demographic and motivational factors that influence concert attendance (Hager & Winkler, 2012; Harlow, 2015; Reynolds, 2017). Research also sought to examine the software other symphonies are utilizing in order fulfill the same purpose. Upon completing this preliminary research, the team identified four different IT solutions to offer the Client and meet the requirements of the organization.

During the first client briefing, the team offered the Client four possible IT solutions that leveraged the power of GIS: Business Analyst Online, Tableau, Google Maps and Bloomerang. After examining the four recommendations, the Client elected to have the team initiate work using Google Maps, as it met the budgetary constraints of the organization and the Client felt comfortable using Google software within the organization.

Software Name	Type/Services	Price
Tableau	<ol style="list-style-type: none"> 1. Desktop Personal: Connect to files like Excel and Google Sheets. 2. Professional: Connect to hundreds of data sources. 3. Online: is the SaaS form of Tableau Server with maintenance, upgrades, and security fully managed by Tableau. Any browser or mobile device. 4. A license, with annual maintenance. 	<p>Tableau Desktop: Personal = \$35 per month. Professional = \$70 per month.</p> <p>Tableau Online: Fully Hosted = \$42 per month.</p>
Business Analyst	<ol style="list-style-type: none"> 1. \$500 per year. 2. Adding \$100 per person. 3. 5 users minimum. 	At least \$1000 per year.
Google Maps	<ol style="list-style-type: none"> 1. Standard Plan: For free and publicly available apps/websites, use the Standard Plan. 2. Free up to 25,000 map loads per day. 3. Premium Plan: If usage limits or require 24/7 technical support and an SLA, contact us for a Premium Plan. 	<p>Standard Plan = FREE</p> <p>Premium = \$0.50 per day.</p>
Bloomerang	<ol style="list-style-type: none"> 1. Donor Search 2. Bloomerang finding email 	\$5,500-\$6,500 first year.

	addresses for your constituents.
3.	0 - 1,000 records.

After receiving approval from the Client, the team began the implementation phase by researching unconventional demographic and motivational factors that have been found to influence concert attendance. Factors such as voter registration and volunteerism were found to have a positive correlation with concert attendance and, consequently, charitable giving. According to Ostrower (2005), “the proportion of respondents registered to vote climbs from 73 percent among non-attendees up to 91 percent among frequent attendees [...] and doing volunteer work rises from 27 to 63 percent” (p. 7). Ostrower (2005) also found that frequent attendees were most likely to be donors. These factors would form the basis for the data collection and visualization in Google Maps for the Client.

The team utilized Google Fusion Tables as a means to generate density maps from different data sources and created a list of the top fifteen potential areas by zip code for the Client (see *Figure 1*).

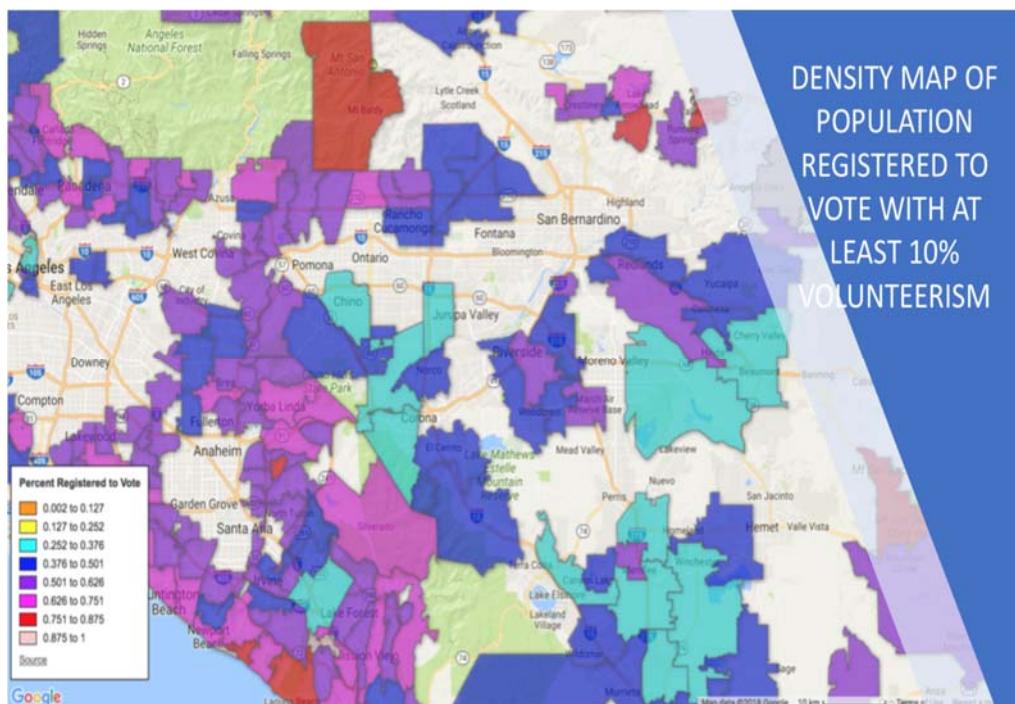


Figure 1. Density map created using Google Fusion Tables displaying zip codes by percent of the population registered to vote and at least 10% volunteerism.

Google Fusion Tables is a web service provided by Google for data management (Google Inc., 2018). These tables are immensely helpful in assisting the user to gather, visualize, and share data available in the organization. Users can merge data from multiple tables and create detailed geographic visualizations from data. As the standard plan for the Google Maps is free, it makes Google Maps the ultimate option for small organizations that have limited financial resources to spend on data management tools and other IT solutions.

The team also sought to provide the Client with the proper tools that would enable them to adequately visualize the data and identify new geographical areas with higher concentrations of potential concert attendees using publicly available data. The scope of this project also included the delivery of tools and resources, such as a video tutorial and user guide, to aid the Client in utilizing the software.

2.2 Symphony Orchestras and Mobile Apps

The effectiveness of mobile apps for symphony orchestras like the SBSO will be examined through an analysis of symphony orchestra IT expenditures as well as mobile app usage and the demographics for mobile app users and orchestra audiences. Reasons for concert performance attendance and non-attendance will also be examined in relation to internet and app usage on mobile devices. Effectiveness will be defined as the viability for mobile apps, such as the Dubuque Symphony Orchestra (DSO) app, to be utilized by symphony orchestras for the purposes of engaging audiences. Establishing a viability could suggest the potential of drawing in new audiences through customized mobile apps for symphony orchestras like the Boston Symphony Orchestra (BSO) app. This research entertains the notion of igniting an interest in performing arts and bringing new musicians and audiences to symphony orchestras via mobile app technology.

3. San Bernardino Symphony Orchestra and Exploring the Use of Mobile Applications by Symphony Orchestras Research

Non-profit organizations are increasingly using technology in their operations. By reviewing the mobile apps for both the BSO and DSO, we can examine the potential for other symphonies, such as the SBSO, to also offer a mobile app to its concert-goers. The BSO is like a path for a smaller organization such as the SBSO to have a better future. However, we also examine the DSO due to having a similar revenue and size as the SBSO while also offering a mobile app.

Considering that the user interface of a mobile app enhances the appeal and likeability of an app thus increasing usage and user loyalty. In extension, this contributes to the increased popularity of the non-profit symphony orchestra which translates to increased revenue for the symphony orchestra. Therefore, orchestras should leverage IT as demonstrated by the development of mobile apps in order to improve attendance and also increase their revenue streams.

The mobile app can help you in everything from planning your visit to purchasing tickets, enjoying digital content and direct engagement through social media plus helpful information while at the concert (Apple, 2018), reads the description to the BSO mobile app on the Apple app store. By mixing digital content and ticketing information related to the BSO, the app allows the orchestra to leverage mobile apps to reach out to fans, share exciting news on the orchestra's planned schedule (Apple, 2018), and communicate information related to the orchestra (Apple, 2018). While the SBSO does not have a mobile application, it is important to note that a mobile app would improve the attendance in SBSO events. According to the BSO, its presence on the Internet has significantly allowed it to attract the highest traffic among orchestras in the U.S. (Boston Symphony Orchestra, 2017). In other words, the mobile app creates awareness of the orchestra's events. In addition, the online presence through mobile and web platforms allows the BSO to promote art music through educational activities (Boston Symphony Orchestra, 2017) as well as attract sponsorship that is necessary for continued activities of the

orchestra. Similarly, there are various mobile apps that could help the symphonies in attaining some new donors like Givelify—this is a free mobile app with non-profit donations with one account, all in one place. This sets up the organization's custom fundraising campaigns and sets donation goals so that the donors can decide where exactly they want their money to go (Givelify, 2018). As it is a free mobile app, every non-profit organization can get benefits from the app. Indianapolis-based Givelify LLC developed and marketed a mobile donation app, is expanding its headquarters and growing its profit largely by the year 2020 (IBJ, 2017). Effectiveness will be defined as the growth for the mobile applications like Givelify to be utilized by the symphony orchestra and music enthusiasts to engage more and more concert-goers and for supporting the arts as well. Moreover, the symphony orchestras can design their mobile apps based on the offerings they are going to present it to their concert-goers and donors. Apps for some orchestras like the Houston Symphony and Buffalo Philharmonic offer only short audio excerpts, while the London Philharmonic offers full works at time and brief excerpts at other times (Tedeschi, 2011).

In addition to looking at current symphony orchestra mobile apps, an analysis will be conducted using various secondary datasets from different sources, including the National Endowment for the Arts and the National Opinion Research Center (NORC). The Survey of Public Participation in the Arts (SPPA) is a rich repository for performing arts participation, behaviors, and demographics offered by the National Endowment for the Arts.

In conjunction, patterns in performing arts participation, motivations, and mobile app/Internet use will be analyzed using data gathered from the General Social Survey (GSS). The GSS is a project of the NORC at the University of Chicago and is among the most widely analyzed sources of information. The GSS features data on various societal attitudes, behaviors, and attributes (Smith, Marsden, Hout, & Kim, 2016). The data that will be used in the analysis includes responses from those who attended a performance in the last 12 months or wanted to attend a performance in the last 12 months but did not

along with the factors that influenced their attendance. In addition, Internet/app usage on mobile devices for this population will be included for analysis.

Lastly, mobile app user behavior will be investigated in further detail through an analysis of the “Worldwide Mobile App User Behavior Dataset” (WMAUBD) (Lim, 2014). This dataset examines mobile app usage and motivations through a survey of 10,208 respondents across 15 countries. In order to make equitable comparisons, only U.S. data from the WMAUB dataset will be used. The survey also considers demographic information and Big Five personality traits of the respondents. An analysis of these three datasets will offer a greater understanding of current mobile app trends among various age groups. In turn, attitudes and behaviors toward the performing arts can be analyzed among the same age groups to determine the viability of symphony orchestras utilizing mobile apps to foster effective audience engagement.

3.1 Symphony Orchestras and IT Expenditures

The team has analyzed three symphony orchestras (BSO, SBSO and DSO) based on their expenditure in IT as reported in the IRS Form 990: Return of Organization Exempt from Income Tax. As compared to SBSO and DSO, BSO addresses a larger audience and their total annual expenses are far more than the two mid-sized symphony orchestras. Out of the three symphonies, BSO and DSO have launched their mobile apps and SBSO does not possess any mobile app yet. As Figure 2 demonstrates the percentage spent on information technology, it is clear from the figure that BSO and SBSO are spending almost the same percentage of their expenses in IT. Even when the percentage of expenses spent on IT is the same, SBSO is not able to increase its audience as compared to BSO. In the year 2014, BSO spent 1.63% of total expenses and SBSO spent 1.87% of their total expenses in IT. SBSO's percentage was greater than BSO in that particular year. SBSO is spending nearly the same expenses as of BSO but they are not able to increase the audience in their concerts.

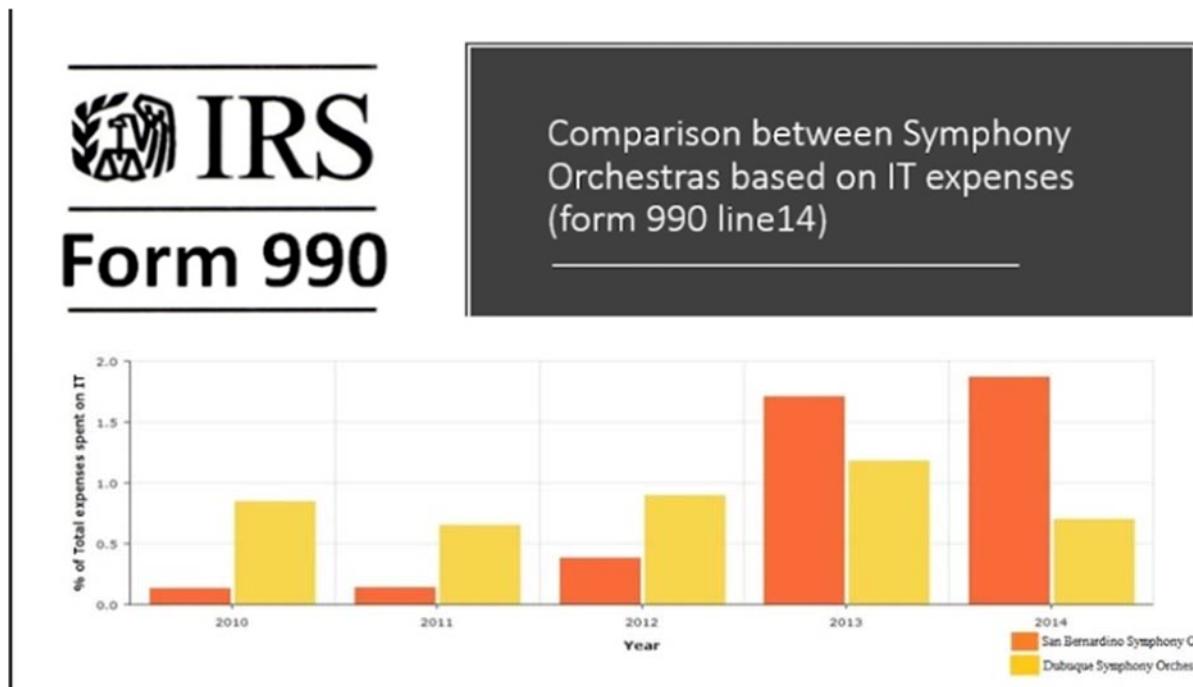


Figure 2: Comparison between symphony orchestras based on IT expenses.

The tax filings in 2015 show that the DSO had revenue of \$1,076,359 with program services contributing 39.9% of the revenue compared to SBSO's revenue of \$635,615 of which 32.1% was raised from program services. In addition, DSO recorded an income of \$102,766 in 2015 and \$1,096,018 in 2014. On the other hand, SBSO managed an income of -\$67,066 and -\$26,039 in 2015 and 2014 respectively. In 2013 and 2012, SBSO had an income of -\$79,550 and -\$62,651 compared to DSO's \$77,860 and \$188,758 respectively.

As SBSO is spending a significant proportion of their expenses on IT, they need to target on mobile apps and get prepared to launch one. It is clear that technology has already significantly changed the way people make, access and consume music. Online streaming services such as Spotify have completely disrupted the industry. In order to improve the customer experience and better engage audiences before, during, and after a performance, SBSO should launch its mobile app which would help them improve audiences and increase their potential donors.

3.2 Concert Attendance and Mobile App Users

This analysis begins with responses from the General Social Survey (Smith et. al, 2016). Barriers for concert non-attendance as well as reasons for concert attendance among two sample populations, mobile app users and mobile app non-users, will be examined. These measures will be converted into proportions due to variance in both sample sizes. Reasons that are statistically significant can provide insight as to the tendencies of mobile app users and mobile app non-users with regard to concert attendance behaviors and attitudes and may suggest a course of action for symphony orchestras as well as assess the viability of mobile app usage for patron engagement.

Figure 3 displays the proportion of mobile app users and mobile app non-users that reported wanting to attend a performance within the 12 months prior to taking the survey but did not.

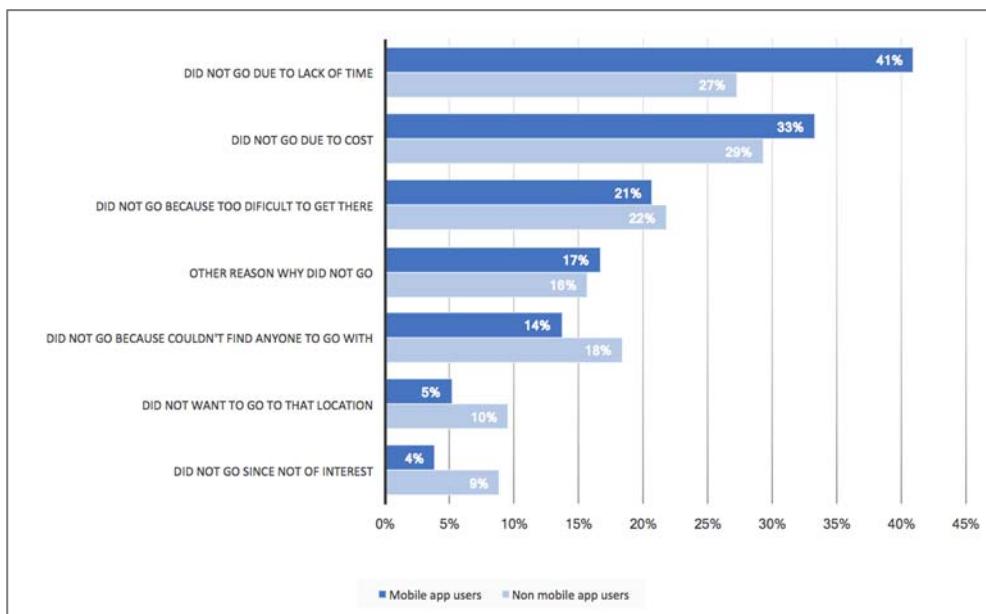


Figure 3:

The reason with the greatest disparity between the two populations, and the most popular reason for mobile app users not attending, was due to lack of time. Our null hypothesis is that there is no significant difference between the reasons why mobile app users and non- users did not attend a performance despite wanting to.

A two-sample z-test between proportions was performed to determine whether there was a significant difference between mobile app users and non-users who did not attend due to lack of time. A significantly greater proportion of mobile app users ($\alpha = 0.01$) wanted to attend a performance but did not due to lack of time, $z(614)=3.057$, $p=.002$. The most popular reason among mobile app non-users for not attending was cost. We found that proportions between mobile app users and non-users that wanted to attend a performance but did not due to cost was not statistically significant, $z(614)=-0.924$, $p=0.356$. Mobile app users were 14% more likely to not attend a performance despite wanting to due to a lack of time. These results identify an area of focus for symphony orchestras seeking to implement a mobile app for prospective concert-goers. How can symphony orchestras appeal to their mobile app audiences knowing approximately 41% do not attend due to lack of time despite wanting to?

Next, we examine major reasons that affected a respondent's decision to attend a performance in the last 12 months. Again, we are examining two populations, mobile app users and mobile app non-users. Proportions will be measured due to variance in sample sizes. (Figure 4) reflects the proportion of respondents that reported the reasons listed as "major reasons" that impacted their decision to attend.

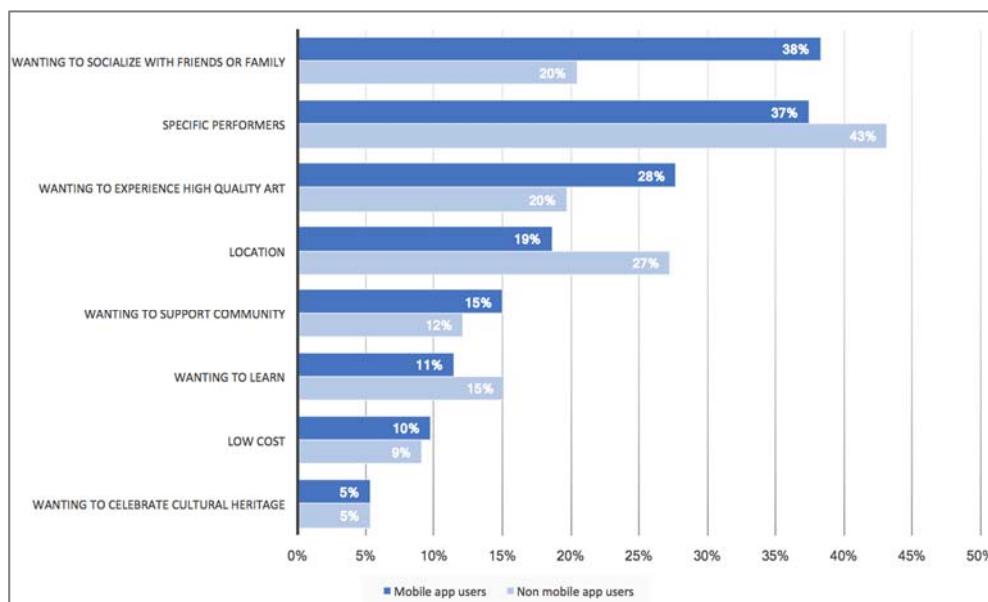


Figure 4: Respondent went to a performance in the last 12 months. Importance of each reason in decision to attend performance. This figure illustrates reasons reported that had a major impact on their decision to attend the performance.

Among mobile app users, the most popular reason for attending was wanting to socialize with friends or family. Among mobile app non-users, the greatest reason that impacted their decision was the performance itself. The difference between mobile app users and non-users was found to be statistically significant ($\alpha = 0.01$) between those who reported wanting to socialize with friends and family as a major reason for attending. Mobile app users were 18% more likely to report socializing as a major reason that impacted their decision to attend a concert performance. Conversely, there is no significant difference between mobile app users and non-users that reported the performance itself as a major reason for attendance despite it being the most popular reason among mobile app non-users.

These findings allude to past research suggesting motivational factors among younger audiences include the desire to socialize (Hager & Winkler, 2012). This assumes mobile app users are primarily comprised of younger audiences. Without assuming age, symphony orchestras can benefit from knowing that prospective attendees who use mobile apps, suggesting they would utilize a symphony app, attend primarily for the social aspect. Symphony orchestras like the SBSO can implement a mobile app and include a socializing aspect in its design. Although age is not reported in this survey, the next comparison will consider age differences and similarities among mobile device users and classical concert attendance, and mobile app user behaviors and attitudes. We will assume that mobile device users will have an inclination toward mobile app usage.

Next, we examine the SPPA and the WMAUB dataset (National Endowment for the Arts, 2012; Lim, 2014). To test the assumption that mobile device users will have an inclination toward mobile app usage, we compare the mean ages in both sample populations using a two-sample t-test. Our null hypothesis is that there will be no significant age differences between mobile device users and mobile app users, irrespective of whether they attended a concert performance. Our results indicate that there are no

significant differences ($\alpha = 0.05$) between average mobile app and mobile device user ages, $t(2750)=1.843$, $p=0.066$. Figures E3 and E4 illustrate the proportion of mobile app users by age from the WMAUB and the proportion of mobile device users and non-users from the SPPA that attended a live classical performance during the last 12 months, respectively.

The greatest proportions of mobile app users are represented in the 25-34 (24%) and 45-54 (21%) age groups as illustrated. The greatest proportions of mobile device users that attended a live classical concert are represented by the 25-34 age group (20%) and 55-64 (19%) age groups. The age group with the third highest proportion of mobile device users that attended a performance is 45-54 (18%). Given that the 45-54 age group is the second highest in mobile app usage and third highest in both mobile device usage and concert attendance, we can suggest that this age group may be more inclined to download a symphony orchestra app and predict that approximately 18% will attend at least one performance.

It is worth noting that the proportion of concert attendees increases with each age group for mobile device non-users. Seventy-five percent of mobile device non-users that attended a concert performance are age 55+. This population is viewed as traditional concert attendees; however, in order for symphony orchestras to survive, they will need to appeal to younger audiences. Mobile apps have the potential to increase engagement by appealing to ages 25-54 who have the propensity to utilize mobile devices and mobile apps as well as attend concert performances. This age range represents more than half (52%) of mobile device users that have attended a concert performance; therefore, we can predict that approximately 52% of people ages 25-54 that have a proclivity toward concert attendance, will also be inclined to use a mobile app.

provides a breakdown of mobile app user preferences from the WMAUB with regard to mobile app downloads. These reasons are considered when deciding whether a user will or will not download a mobile app. Symphony orchestra can gain valuable insight as to what they should consider when creating and implementing a mobile app.

4. Findings and Observations

A mobile application offers a way and platform through which an orchestra is able to connect with its audience and disseminate this information to its followers. Effectively, this is a model through which an orchestra like SBSO is able to create a channel of direct engagement with its audience and provide them with digital content through the means of this platform. For instance, the BSO has been able to use its mobile app in creating an online conversation, creating a program schedule and enabling users to buy tickets for its events. Through the mobile app, the platform is able to create and disseminate digital content in the form of its audios which users can listen, share and download for local use. In addition, users can also create a playlist of their favorite music tunes. Consequently, the mobile app has enabled the orchestra to establish an ardent followership and a group of fans that resonate closely with the music and content that the orchestra offers (Apple, 2018).

In addition, the benefits of a mobile app are not just limited to its ability to create an ardent following, but also in enabling users to create their own music according to their taste through the utilities that it provides. For instance, a mobile app that has been created by Cadenza enables users to create music which resonates with musical possibilities as manifested in the course of expression. "The rich harmony and tapestry that it creates goes a long way in augmenting the process of music making and establishing a give and take music experience that closely resembles the course of typical conversation" (Apple, 2018). In this process of music making, the user has the ability to record themselves and equally create content which they can listen to themselves or share with others through this same platform.

In the process, the use of a mobile app also enabled the musical symphony like SBSO to generate new ideas about ways through which to improve their performance and music. In the process, they are able to improve their musical confidence as part of their delivery process. Further and even more important is the role that a mobile app plays in incorporating the benefits of artificial intelligence into the music making process. Through the mobile app and its utilities, it is possible to create adjustments in the playing so as to fit the content features that are part of the recording. For instance, the use of Cadenza

enables the user to fit the playing instrument with in terms of tune tempo and orchestral performance by allowing the mobile app to listen and simulate the rest of these components in the process (Sona Cadenza, 2016). The mobile app thus enables users to discover and enjoy music as resonates with their heart, a component that is crucial, be it is the course of personal enjoyment or during the course of addition, thereby making the entire process relatively simpler, elaborate and more meaningful.

It is important to note that the BSO's mobile app allows for user review and criticism of the mobile app. The user review section allows the orchestra to improve the mobile app with regards to the consumer complaints (Apple, 2018). The involvement of users of the mobile apps also enhances brand image of the orchestras to the extent that it catalyzes interactive technology with regards to entertainment and educational values of the mobile apps (Palumbo, Dominici, & Basile, 2013). Users tend to identify with brands that enrich their life. In addition, the videos accessible on the mobile app act as a pull factor for converting online users of the mobile app to physical visitors to paid concerts.

After researching and using the IRS Form 990 for the non-profit organizations, many organizations profited. DSO was the closest in terms of revenue and expenses for IT and it may vary in each year but in a close range to compare. Where it was the most differences after the year 2011, where the SBSO drop in each year (Nonprofit Explorer, 2017), while DSO was improving until it reached its peak in 2014-2015, the same years that they doubled the expenses of IT (Nonprofit Explorer, 2017), showing us the importance of the IT and mobile apps. While the revenue for SBSO in the year 2016 was \$493,186 (Nonprofit Explorer, 2017), the DSO generated \$896,132 with an operational deficit of - \$152,069 and -\$89,241 (Nonprofit Explorer, 2017) respectively. The tax filings in 2015 show that DSO had revenue of \$1,076,359 with program services contributing 39.9% (Nonprofit Explorer, 2017) of the revenue compared to SBSO's revenue of \$635,615 of which 32.1% was raised from program services (Nonprofit Explorer, 2017). In addition, DSO recorded an income of \$102,766 in 2015 and \$1,096,018 in 2014 (Nonprofit Explorer, 2017). On the other hand, SBSO managed an income of -\$67,066 and -\$26,039 in 2015 and 2014 respectively (Nonprofit Explorer, 2017). In 2013 and 2012, SBSO had an income of -

\$79,550 and -\$62,651 (Nonprofit Explorer, 2017) compared to DSO's \$77,860 and \$188,758 (Nonprofit Explorer, 2017) respectively.

Arguably, the ability of DSO to finance its expenses effectively is attributable to its absorption of technology in its activities. The rapid expansion of mobile-based application technologies allows a symphony orchestra to enhance its brand visibility and brand competitiveness. Promoting real time access news and updates on the DSO, the DSO mobile app is organized into several windows (AppShopper, 2018). The section on events allows one to gain useful information related to performance of orchestral repertoire with regards to dates and venues (AppShopper, 2018).

The free mobile app also allows fans of the DSO to enjoy video performances (AppShopper, 2018) of the DSO. Additionally, the section dubbed buzz makes it possible for individuals to follow the news (AppShopper, 2018) related to the symphony orchestra. The detailed information on events and news on the DSO mobile app allows individuals to obtain information related to performances leading to a high turnout during programs (Boice, 2014). Such a higher turnout during program services leads to increased generation of revenue through ticketing and patronage. Additionally, the integration of the app with social networking sites allows people to create awareness of the Orchestra's programs among people who do not use the app further enhancing the turnout and eventual fees collected through tickets.

5. Recommendations

The possibilities in the use of a mobile app for an orchestra are astounding. At some level, the mobile app and its utilities have the ability to deliver effectively and produce musical content which resonates closely and mirrors the output in a real orchestra. In this respect, the innate abilities of an orchestra mobile application and its cutting-edge algorithms would be beneficial and crucial in analyzing and generating content to the desires and intricate requirements that would be part of the orchestra (Sona Cadenza, 2016). The mobile app utility would contribute in creating new music and enabling the users to

augment and leverage on their ability through an improvement in utility that this technology provides. For these reasons, there is much that SBSO stands to gain from a mobile app for its orchestra.

Nonprofit organizations should leverage technology to expand generation of revenue. The organizations should design user intuitive mobile apps to enhance user experience and maximize educational and entertainment utility of their apps. Apps that enrich user experience often lead to increased attendance during events organized by the non-profits. Improved attendance during events increases revenue thus allowing organizations to cover their running expenses. Overall, non-profit organizations should leverage technology to increase attendance and revenue.

6. Strategic IT Plan for SBSO

The current state in technology is highly dynamic. For an orchestra that has evolved from its inaugural concert back in 1887 and now in its 140th season, the role of technological change provides the most solid avenue of establishing an IT strategy going forward. Currently, the Orchestra has grown to a level where it reaches its audience through radio, television and over the internet (Boston Symphony Orchestra, 2017). Effectively, the most promising portal in the current age lies in the use of internet and mobile phones as a way of further spurring its growth and spreading out its ubiquity. A suggested IT plan would involve creating a mobile app for a symphony that enables the user to access the physical features in the live symphony through the touch of their hands. The mobile platform as an end provides the most promising opportunity of growth and spread that would capitalize on the large number of users in the United States and indeed the rest of the globe (Statista, 2017).

6.1 Ethical IT Skills in Managing and Securing Client Data for the Mobile App

The new mobile app will be developed and used under the secure end to end encryption model which aims to protect user data. Credit information can only be accessed by the owner and through the role of encryption, it will be impossible to cut through and potentially expose

vulnerable user information to third parties. Additionally, the app will operate under the third-party development agreements that are currently in use with both Apple and Android where the user is entirely entitled to the information (Boice, 2014). Under this arrangement, the operating system ownership only acts as the carrier while the owner reserves the right to all user information. Equally, the current technology space is highly secure, and the basic role of encryption is by itself a fool proof mechanism that would ensure that user and donor data is managed and secured in a highly effective manner.

6.2 Technological Dimensions and Decision-Making Frameworks

The main aim behind this development lies in the need to provide value and convenience to the user at the comfort of their palms. The main utility that would have been incorporated in this case include enhanced visibility of the brand and the ability to heighten their competitive acumen in the process. For this purpose, the mobile app will include an ability for the users to exercise autonomy through easy to decipher instructions and greater control over its operations and function. Beyond the security consideration that is meant to ensure the safety of user data, the other components will be largely left under the control of the individual user (Business of Apps, 2017).

6.3 Risk and Return Information for a Mobile App

As a nonprofit institution, SBSO depends on direct donations in funding over 70 percent of its operations. The role of a mobile app as part of its IT initiatives would serve to optimize its operations and spur its revenues by enabling the entity to harness on the direct benefits that accrue from developing mobile apps. At a glimpse, a new app will enable SBSO to enhance its direct revenue generation, user acquisition and provide additional utilities in the aim of increasing the value in the established experience which is being given to the audience. A look at the risk and return elements further illustrates that the feasibility of this initiative is very high. Beyond the initial development costs, the app project will incur zero additional costs in customer acquisition (Boston Symphony Orchestra, 2017). Equally, the costs of maintenance are projected to be significantly low. Organically speaking, the Orchestra would have

created a full fletched performance line without making a single dime of investment into the process of marketing. Regarding return, the firm under its current operating model is bound to reap maximum benefits through the millions of users that would access its service through the mobile app platform and consequently work as a vibrant user base and source of direct donations towards its operations.

6.4 Business Requirements and Formulated Technology Solutions

The operating framework for the establishment of a mobile app for the Symphony would basically focus on the user demographic components in addition to the safety and operating considerations from the provider's end. For the app development process, the firm will have to consider the initial cost of setting up and rolling out the app through the different carriers, mainly Apple and Android. Subsequently, the maintenance component would involve taking care of security, both for the user and the firm as the provider of this solution from third party intrusion (Business of Apps, 2017).

7. Conclusion

Creativity thrives when people work together on a team. Collaborating on a project creates an enthusiasm for leaning and also maximizes the shared knowledge and helps the team members learn new skills (Mattson, 2015). In this project, our team members have co-operated with each other during all the stages of the project. The team had put in all the efforts to find the required data and we were successful too. All the team members were friendly and delivered their assigned tasks on time. Small teams have various challenges too any and many of the problems are similar to every team face- disagreements, unclear priorities (Harrin, 2016). It was a successful and educative journey delivering the project deliverables and meeting the course requirements on proper time with appropriate material. All of us were very communicative with each other to take advice or ideas of other members. All of the team members have devoted immense effort in the project and thus we have made our project a success. Without the right team in place, any strategy and plan has the potential of completely falling apart (Palmer, 2016).

The project team was able to successfully offer an IT solution that met the needs and requirements of the Client. Moreover, in our research on the effectiveness of mobile app usage for improved concert attendance and audience engagement, we evaluated three symphony orchestras in relation to proportion of IT expenditures and revenues. Additionally, we examined age in relation to concert attendance and mobile app usage and found higher propensities among certain age groups to use mobile apps as well as attend an orchestral performance. Our findings suggest potential for the use of mobile apps by symphony orchestras as a means of engaging their audiences and suggests focusing on the 25-34 and 55-64 age groups. The BSO sets the example for aspiring orchestras looking to grow in size and revenue. Further areas of study can include other areas of IT such as customer relationship management (CRM) and business intelligence (BI) and analytics tools. Most importantly, we hope shining a light on one segment of the performing arts will invite others to engage in the necessary conversation about the significance of the arts in our society and in education. Only then can we begin to bridge the gap between science and art and elevate the performing arts once again, this time leveraging the power of technology.

References

- Apple. (2018). Boston Symphony Orchestra. Retrieved from App Store Preview:
<https://itunes.apple.com/us/app/boston-symphony-orchestra/id1032626688?mt=8>
- AppShopper. (2018). Dubuque Symphony Orchestra. Retrieved from AppShopper:
<http://appshopper.com/music/dubuque-symphony-orchestra>
- Arts Index. (2016). 2016 National Arts Index: An annual measure of the vitality of arts and culture in the United States: 2002-2013. Retrieved from <http://www.artsindexusa.org/2016-national-arts-index>
- Baumol, W. J. & Bowen, W. G. (1965). On the performing arts: The anatomy of their economic problems. *The American Economic Review*, 55(1), pp. 495-502.

Bloomerang (2017). Our pricing is simple and straightforward. Retrieved from

<https://bloomerang.co/pricing>

Boice, D. (2014, Apr 22). Going mobile: The importance of mobile apps for nonprofits. Retrieved from Nonprofit Technology Network: <https://www.nten.org/article/going-mobile-the-importance-of-mobile-apps-for-nonprofits/>

Boston Symphony Orchestra. (2017). The History of the BSO. Retrieved from Boston Symphony Orchestra: <https://www.bso.org/brands/bso/about-us/historyarchives/the-history-of-the-bso.aspx>

Business of Apps. (2017). Mobile Apps Benefits. Retrieved from <http://www.businessofapps.com/mobile-app-benefits/>

Givelify. (2018). Why take your non-profit fundraising mobile? Retrieved from: <https://www.givelify.com/nonprofits/>

Google Inc. (2018). Google fusion tables REST API. Retrieved from

https://developers.google.com/fusiontables/docs/v2/getting_started

Hager, M. A. & Winkler, M. K. (2012). Motivational and demographic factors for performing arts attendance across place and form. *Nonprofit and Voluntary Sector Quarterly* 41(3), pp. 474-496.

Harlow, B. (2015). Taking out the guesswork: Using research to build arts audiences. *The Wallace Foundation*. Retrieved from <http://www.wallacefoundation.org/knowledge-center/pages/default.aspx>

Harrin, E. (2016). 5 common problems for small project teams. Retrieved from

<https://www.liquidplanner.com/blog/5-common-problems-small-project-teams/>

IBJ Staff (2017, Jan 26). Donation-app firm Givelify plans 40-job expansion downtown. *IBJ*. Retrieved from <https://www.ibj.com/articles/62218-donation-app-firm-givelify-plans-40-job-expansion-downtown>

Internal Revenue Service. (2016). Return of organization exempt from income tax: San

Bernardino Symphony Association [Form 990(c)(3)]. Retrieved from

<https://projects.propublica.org/nonprofits/organizations/956153923>

Intriligator, W. (2018). Dubuque Symphony Orchestra mobile app. Retrieved from <http://www.dubuquesymphony.org/media-gallery/dubuque-symphony-orchestra-mobile-app>

Leka, O. (2017). Database of android apps [Data file]. Available from

<https://www.kaggle.com/orgesleka/android-apps/data>

Lim, S. L. (2014). Worldwide mobile app user behavior dataset (1.0) [Data file and code book].

Available from Harvard Dataverse Web site:
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/27459&version=1.0>

Mattson, D. (2015). 6 benefits of teamwork in the workplace. *Sandler Training*.

Retrieved from <https://www.sandler.com/blog/6-benefits-of-teamwork-in-the-workplace>

National Endowment for the Arts. (2012). *2012 Survey of Public Participation in the Arts* [Data file and code book]. Available from the National Archive of Data on Arts and Culture:

<https://www.icpsr.umich.edu/icpsrweb/NADAC/studies/35168#cite>

Nonprofit Explorer (2017). Boston Symphony Orchestra Inc. Retrieved from Nonprofit Explorer: <https://projects.propublica.org/nonprofits/organizations/42103550>

Nonprofit Explorer (2017). Dubuque Symphony Orchestra. Retrieved from Nonprofit Explorer:

<https://projects.propublica.org/nonprofits/organizations/237429727>

Nonprofit Explorer (2017). San Bernardino Symphony Orchestra. Retrieved from Nonprofit Explorer:

<https://projects.propublica.org/nonprofits/organizations/956153923>

Ostrower, F. (2005). The diversity of cultural participation: Findings from a national survey. *The*

Urban Institute. Retrieved from <http://www.wallacefoundation.org/knowledge-center/pages/default.aspx>

Palmer, E. (2016). Five factors that lead to successful projects. Retrieved from

<https://project-management.com/five-factors-that-lead-to-successful-projects/>

Palumbo, F., Dominici, G., & Basile, G. (2013). Designing a mobile app for museums according to the

drivers of visitor satisfaction. Retrieved from www.wseas.us/e-library/conferences/2013/Dubrovnik/MATREFC/MATREFC-24.pdf

Peppler, K. (2013). New opportunities for interest-driven arts learning in a digital age. *The*

Wallace Foundation. Retrieved from <http://www.wallacefoundation.org/knowledge-center/Documents/New-Opportunities-for-Interest-Driven-Arts-Learning-in-a-Digital-Age.pdf>

Reynolds, J. (2017). #Hashtag orchestra. *Symphony Magazine*, 68(1), 24-30.

Rosen, J. (2017). Vision of orchestras. *Symphony Magazine*. pp. 20-24. Retrieved from

https://americanorchestras.org/images/stories/symphony_magazine/summer2017/Visions%20of%20Orchestras%20-%20Critical%20Questions.pdf

San Bernardino Symphony Orchestra. (2017). About us.

Retrieved from <http://www.sanbernardinosymphony.org/index.html>

Smith, T. W., Marsden, P., Hout, M., and Kim, J. (2016). General Social Surveys [Data file].

Available from the NORC GSS Data Explorer Web site: gssdataexplorer.norc.org.

Sona Cadenza. (2016). Cadenza. Retrieved from <http://www.sonacadenza.com/the-app/>

Statista. (2017). Number of smartphone users in the United States from 2010 to 2022. Retrieved from
<https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/>

Tableau (2017). Pricing. Retrieved from <https://www.tableau.com/pricing>

Tedeschi, B. (2011, Jul 27). To fill a gap in commercial radio, classically trained apps. *The New York Times*. Retrieved from <http://www.nytimes.com/2011/07/28/technology/personaltech/classical-music-apps-for-smartphones.html>.

Improved Text mining for bulk data using Deep learning approach

Indumathi A
PG Scholar,

Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Perumal P
Professor,

Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Abstract- Text document clustering and similarity detection is the major part of document management, where every document should be identified by its key terms and domain knowledge. Based on the similarity, the documents are grouped into clusters. For document similarity calculation there are several approaches were proposed in the existing system. But the existing system is either term based or pattern based. And those systems suffered from several problems. To make a revolution in this challenging environment, the proposed system presents an innovative model for document similarity by applying back propagation time stamp algorithm. It discovers patterns in text documents as higher level features and creates a network for fast grouping. It also detects the most appropriate patterns based on its weight and BPTT performs the document similarity measures. Using this approach, the document can be categorized easily. In order to perform the above, a new approach is used. This helps to reduce the training process problems. The above framework is named as BPTT. The BPTT has implemented and evaluated using dot net platform with different set of datasets.

1. INTRODUCTION

The capacity of storage data becomes huge amount of the technology of computer hardware develops. So amount of data is increasing exponentially, the information required by the users become varies. Actually users deal with textual data more than the numerical data. It is very difficult to apply techniques of data mining to textual data instead of numerical data. Text mining [1] is finding interesting regularities in large Textual datasets. The text mining studies are gaining more importance recently because of the availability of the increasing number of the documents from a variety of sources. Which include unstructured and semi structured information. The main functions [2] of the text mining include text summarization, text categorization and text clustering. The Text of this paper is restricted to text categorization.

“Text mining” is increasingly being used to denote all the tasks that, by analyzing large quantities of text and detecting usage patterns, try to extract probably useful (although only probably correct) information.

nonlinear modules that methods transforms the representation at one level (starting with the raw input) into a higher representation slightly more abstract level, with the composition of enough such transformations, and very complex functions can be learned. Deep learning approach of learning algorithm, feature extraction can improve the accuracy of learning algorithm and shorten the time. Selection from the document each part can reflect the information on the text classification, and the calculation of weight is called the text feature extraction.

2. RELATED WORK

In the recent years, the progress of web and social network technologies have led to a massive interest in the classification of text documents containing links or other meta-information and many studies on classification algorithms have been done by many researches. In this section we will do a review to these works and show the focus points of them. As we will see, the novelty of our work is appears by studying almost all the modification and improvements to each algorithm. Focused [4] on specific changes which are applicable for the text classification. They used, as text classification algorithms, Decision Trees, Pattern (Rule) based Classifiers, SVM Classifiers, Neural Network Classifiers, Bayesian (Generative) Classifiers, nearest neighbor classifiers, and genetic algorithm based classifier. They are discussed the methods used for in text classification and described these methods for text classification. To text classification [5] process of text classification as well as the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, principal and performance. The theory and methods of text classification and text mining, the important

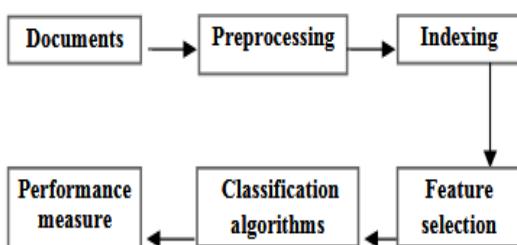


Fig.1.1 Document classification process

Deep learning approach [3] are representation learning methods with multiple levels of representation, but

algorithms that are text classification. In features [6] of each category by using the information. In this performance for this algorithm was reasonable where they showed that feature selection in the decision tree algorithm was particle effective in dealing with the large feature sets common in text categorization. They used the feature extraction and modified the used algorithm. They are many improvements to the well-known algorithms for text classification. The improvements in algorithm can be modification/addition to the algorithm and the learner.

3. PROPOSED SYSTEM

In this proposed method derives text similarity from semantic and syntactic information contained in the similarities text. A text is considered to be a sequence of words each of which carries useful information. The words along with their combination structure make a text convey a unique meaning.

Clustering is the most widely used technique in text mining process. It organizes a large quantity of disordered text documents into a small number of meaningful and sticking together clusters, they provides the foundation for something for intuitive and informative navigation and browsing mechanisms. Text-clustering is to divide a collection of text- documents into several categories so that documents in the same concept describe that identical topic such as classical music. Text Clustering efficiently groups documents with similar collection into same cluster. Similarity between objects is measured within the use of similarity function.

The back propagation based Time algorithm is used for fast document similarity analysis. In a recurrent neural network, errors can be propagated further, i.e. more than 2 layers, in order to capture longer history information. This process is usually called unfolding. The recurrent weight in an unfolded RNN is duplicated spatially for an arbitrary number of time steps, here referred to as τ . In accordance with Equation 1, errors are thus propagated backward as:

$$\delta_{pj}(t-1) = \sum_h^m \delta_{ph}(t) u_{hj} f'(s_{pj}(t-1))$$

Where,

h is the index of hidden node at time t . The ignorance deltas of higher layer weights are calculated recursively. After obtaining all the error deltas, weights are folded back adding up to one big change for each unfolded weights.

Figure 3.1 shows the classification using similarity. The proposed algorithm consists of two stages; the first stage is clustering, and the second stage is flow level classification.

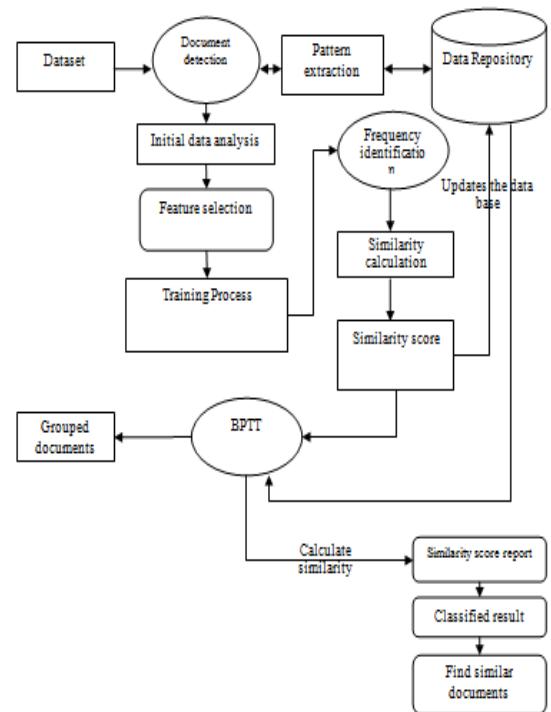


Figure 3.1 Flow of BPTT

The back propagation training algorithm similar documents from the big data environment. The mathematical method used to calculate derivatives of chain rule. This is a training algorithm for updating network weights to decrease error.

The BPTT has the following steps.

- The pattern of input and propagate it through time to get on output
- Analyze the predicted outputs to the expected outputs and calculate the error.
- Calculate the derivative weights of the error.
- Adjust the weights to minimize the error.
- Repeat.

4. DATA SET

The proposed system used real-time and synthetic datasets. Different corpus adopts different rules and models. Some have documents with specialized vocabulary containing words that are repeated frequently. On the other hand, corpus derived from certain sources exhibit creative writing style with word occurrences seldom repeated in their documents. Further details, including discussion of previous versions of the collection (e.g. Reuters-22173), are available in the website. The dataset is available <http://www.research.att.com/~lewis/reuters21578.html> and <ftp://canberra.cs.umass.edu/pub/reuters>. It has 90 specialized categories. All the 90 categories can be used in the experiments.

5. RESULT AND DISCUSSION

Assessment of overall performance: In this subsection, the report gives the results and overall performance of the proposed BPTT model. So, the first process is comparing its accuracy with that obtained by BPTT. Then the next illustrate the variety of prior solutions present in the final iteration. Finally this gives the salient performance parameters of the best pattern obtained for each dataset and compare them with previously reported results.

Comparison with BPTT Model: In order to compare with the BPTT approach with a existing system, this chapter conducted BPTT based document grouping process using patterns of each document in each of the data sets. The existing BPTT was trained and tested for each corpus. Table 5.1 tabulates the accuracy results obtained for the two approaches.

- a) The proposed collaborative approach performs comparatively better than BPTT for both datasets of each corpus. The average accuracy for the collaborative method is 95.55% as compared with 81.44% with the BPTT method, thus giving an improvement of 25%.

Table 5.1 Performance Comparison between existing and BPTT approaches.

Datasets	Accuracy using existing system (%)	Accuracy using BPTT (%)
R21578	86	96.5
Dataset1	84	97
Dataset2	83	96

- b) In cases where the BPTT method gave acceptable results, i.e. 86% for the R21578 dataset and 84.5 % for the Dataset1, the approach enhanced it in both cases to 96.5% and 97% respectively.
- c) For the synthetic and large dataset 2, the BPTT approach led to rather poor results which were dramatically improved with a collaborative approach. For instance, the classification accuracy of the Dataset2 was only 83% using existing approach. This improved to as much as 96% with the BPTT approach. This is because the DC system

was able to utilize the context based pattern maximally in the domain corpus.

Nowadays, document classification in system requires high detection rate and low false alarm rate, thus the research compares accuracy, detection rate and false alarm rate, and lists the comparison results of various documents.

Table: 5.2 Performance comparison table.

Metrics	Existing	Proposed
Similarity calculation	4.3	2.2
Time(ms)		
Efficiency	Ordinary	Better
Accuracy (%)	90.7	97.5

The comparison between existing and proposed system based on the Training time. The training time of the other classification algorithms with the proposed system.

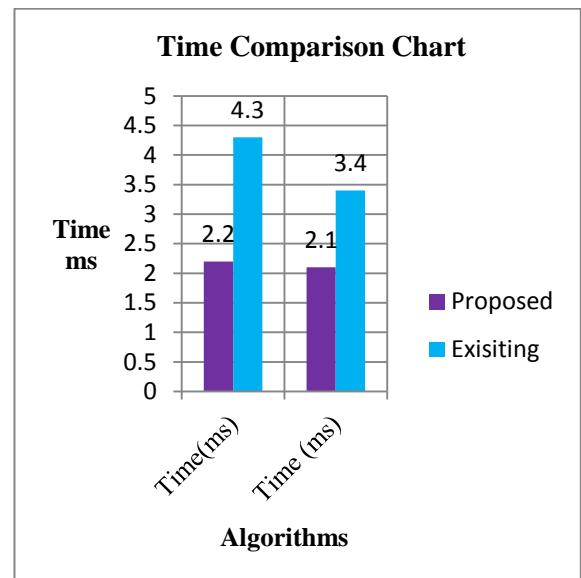


Fig: 5.2 Time comparison between existing cosine similarity and proposed BPTT

6. CONCLUSION

Mining is a significant research area which is gaining an increasing popularity in the recent years. The similarity between the text documents is an important operation of text mining. Text Classification is an important application area in information retrieval, text mining. Because classifying

millions of text document manually is an expensive and time consuming task. In order to reduce the training process, a BPTT approach is implemented in this project. The system proposed an effective method with various patterns for document grouping. This paper also performed the similarity measure for the given two documents based on its external and gathered features.

7. REFERENCES

- [1] Hung Chim and Xiaotie Deng, (2008) “Efficient Phrase-Based Document Similarity for Clustering,” IEEE Transactions on Knowledge and Data Engineering, Vol. 20, Issue. 9, pp. 1217 – 1229.
- [2] Wael H. Gomaa Aly A. Fahmy,(2013) “A Survey of Text Similarity Approaches,” International Journal of Computer Applications, Vol.68, pp.1-13.
- [3] B.Pangand L.Lee, (2008) “Opinion mining and text analysis,” International Conference on Information Technology, Vol.2, Issue.2, pp.1–35.
- [4] Pablo Basanta-Val, Neil C.Audsley, Andy J. Wellings, Ian Gray, and Norberto Fernandez-Garcia, (2016) “Architecting Time-Critical Big-Data Systems,” IEEE Transactions on Big Data, Vol. 2, pp.1- 4.
- [5] Amita Verma, Ashwani kumar, (2014) “Performance Enhancement of K-Means Clustering Algorithms for High Dimensional Data sets,” International Journal of Advanced Research, Vol. 4, Issue. 1, pp 791-796.
- [6] Y.Lu,C.Zhai, and N.Sundaresan, (2009) “Rated aspect summarization Of short comments,” International Conference on World Wide Web, Vol.1, pp.131–140.
- [7] Potts C, (2010) “From frequency to meaning: vector space models of semantics,” Journal Artif Intell, Vol.4, Issue.3 ,pp.1-8.
- [8] K.Fanand, C.H.Chang,(2010) “Text-oriented contextual advertising,” Knowledge and Information Systems, Vol.23, Issue.3, pp. 321–344.
- [9] Manning CD, Raghavan P, Schutze H, (2008) “Introduction to information retrieval,” IEEE Conference on Information Technology, Vol.6, Issues.2, pp. 279–288.
- [10] Wellings AJ, Audsley NC, Basanta-Val P, Fernndez Garca N, (2015) “Improving the predictability of distributed stream processors,” Science Direct on Computer Application, Vol.52, pp. 22–36.
- [11] M.Hu and B.Liu, (2004) “Mining and summarizing customer reviews,” in KDD2004, pp.168–177.
- [12] Kumar S, Toshniwal D (2016), “A novel framework to analyze road accident time series data,” Journal of Big Data, Vol.3, pp.1-8.
- [13] H. Becker, M. Naaman, and L. Gravano,(2010) “Learning similarity metrics for event identification in social media,” The third ACM international conference on Web search and data mining, Macau, China, pp.131-142
- .
- [14] Kumar S, Toshniwal D, (2016) “Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient,” Journal of Big Data, Vol.3(1), pp.1–11.
- [15] Michie MG, (1982) “Use of the bray-curtis similarity measure in cluster analysis of foraminiferal data,” Journal of Big Data, Vol.14, pp.661–667.

A Ranking Model for Software Requirements Prioritization during Requirements Engineering: a case study

Ishaya P. Gambo
Department of Computer
Science and Engineering,
Faculty of Technology
Obafemi Awolowo
University
Ile-Ife, Nigeria
ipgamb@gmail.com

Rhoda N. Ikono
Department of Computer
Science and Engineering,
Faculty of Technology
Obafemi Awolowo
University
Ile-Ife, Nigeria
rhoda_u@yahoo.com

Philip O. Achimugu
Department of Computer
Science, Lead City
University, Ibadan,
Nigeria
check4philo@gmail.com

Olaronke G. Iroju
Department of Computer
Science, Adeyemi College
of Education,
Ondo,Nigeria
rojuolaronke@gmail.com

Abstract- Software requirements prioritization is a recognized practice in requirements engineering (RE) that facilitates the management of stakeholders' subjective views as specified in their requirements listing. Since RE process is naturally collaborative in nature, the intensiveness from both knowledge and human perspectives opens up the problem of decision making on requirements, which can be facilitated by requirements prioritization. However, due to the large volume of requirements elicited when considering an ultra-large-scale system, existing prioritization techniques proposed so far suffer some setbacks in terms of efficiency, effectiveness and scalability. This paper employed the use of a more efficient ranking algorithm for requirements prioritization based on the limitations of existing techniques. The major objective is to provide a well-defined ranking procedure through analysis, suitable for prioritizing software requirements. An empirical evaluation of the proposed technique was made using a typical scenario of the Pharmacy Information System at the Obafemi Awolowo University Teaching Hospital Complex (OAUTHC) as a case study. The results showed the computation of the positive ideal solution (PIS) and negative ideal solution (NIS), as well as the closeness coefficient (CC) for 4 requirements across 3 stakeholders. The CC showed the final ranks of requirements, where R4 with 2.09 point is the most valued requirements, while R1 and R2 with CC of 1.37 and 1.05 were next in the order of priority respectively. The CC provides the medium through which problems of multiple criteria decision making can be handled, so as to determine the order of priority of the available alternatives. The paper conveyed encouraging evidence for the software engineering community that is capable of resolving redundant specified requirements, thereby providing the potential that will facilitate effective and efficient decision making in handling the differences amongst requirements that have been prioritized. Thus, prioritizing software requirements with the recommended ranking procedure during software development is crucial and vital in order to reduce development cost.

Keywords: Software systems, requirements engineering, requirements prioritization, ranking algorithm.

I. INTRODUCTION

In engineering software systems during requirements engineering process, requirements prioritization is essential for the purpose of

implementing an agreeably ultra-large scale system. For instance, in developing critical systems with large number of stakeholders' requirements, prioritization can help in facilitating the choice of the final requirements listing as specified by the stakeholders [1]. Thus, prioritization of elicited requirements is essential in the development of software products that will meet the desired goals of stakeholders. The requirements in this context include useful information that will satisfy the need of the users or project stakeholders [2], and prioritizing them will help to prevent breaches in contracts such as budget over-shoot, exceeding delivery time and missing out important requirements during implementation [3]. We therefore see requirements prioritization as a process of managing the subjective views of stakeholders as specified in their requirements listing. This is with the aim of handling and negotiating the contradictory and conflicting expectations from each stakeholder among other reasons specified in Liaskos et al [4].

However, the selection and prioritization of requirements for the purpose of engendering a system of high quality is seen as a major challenge in software development [5]. It is obvious in literature that prioritization supports the recognition of all the foremost requirements as perceived by relevant stakeholders [6], and it is the activity required for the selection of appropriate requirements [7]. This is with a view of implementing the core sets of requirements with respect to cost, quality, available resources and delivery time [8, 9]. Hence, this paper further buttressed the importance of requirements prioritization and suggests that unmistakable ranking of requirements is an essential success factor for ensuring efficient requirements engineering process. In this case, the importance of every requirement should be based on each stakeholder's subjective view, which makes it multidimensional since it is dependent on the stakeholders' perspective [10]. The primary objective of ranking stakeholders' requirements in a software development process is to aid the analysis of the to-be system by providing the order of their implementation plan amidst available alternatives [11, 12, 13, 14]. However, due to the large volume of data (referred to as stakeholders' requirements), a number of prioritization techniques

proposed so far suffer some setbacks. The resultant effect of such setbacks makes it impossible for software developers to unveil interesting and actionable information about the expected requirements for the software project and product.

In this paper, we propose a well suited ranking procedure given the basic flow of processes expected in the prioritization of software requirements, and the corresponding algorithm. The remaining part of this article has been structured as follows: The second section reveals the related works, the third section deals with the proposed method indicating the various tasks of the ranking procedures and emphasized on its suitability and relevance for software prioritization. The fourth section presents the experimental set-up also covering the empirical evaluation of proposed method on a case study; the fifth section presents and discusses the results; while the sixth section concludes the research and identify suggested future work.

II. RELATED WORKS

In the software engineering literatures, several authors support the need for prioritizing software requirements in making the right choice from the different viewpoint aspects of stakeholders. For instance, the authors in [15, 16, 17, 18, 19] considered how requirements can be prioritized in a multi-team agile context, considering the change-driven nature of the agile methodology. Even in the goal oriented requirements engineering (GORE) methodology, prioritization is considered most important for the purpose of picking out the goals with respect to domain specific needs [20, 21, 22, 23]. However, providing a well-defined ranking procedure suitable for prioritizing software requirements across the various aspects and methodology is essential for implementation plan amidst available alternatives.

Consequently, the advantages of prioritizing software requirements is beneficial to software development project and practises as established for decades in the literature. For example, Pitangueira *et al.* [24] conducted a systematic review to investigate and analyze the various approaches proposed in literature to address software prioritization and selection problems. Their emphasis was based on Search-Based Software Engineering (SBSE), and their findings indicated the aspects addressed by most researchers, and the most prominent techniques used based on the defined problem. Additionally, Gambo *et al.* [1] considered the possibility of integrating Fuzzy Multi Criteria Decision Making (FMCMD) alongside similarity measures and target-based approaches to requirements prioritization using linguistic values of triangular fuzzy numbers. The emphasis in [1] was on how to avert subsequent system failure by making precise and accurate decision in developing large scale software systems. However, an algorithm that will rank and/order these sets of requirements is essential to the enhancement of the proposed techniques in [1].

Other examples of related works from literature include: (1) the work of Babar *et al.* [5] on the analysis of various issues associated with existing prioritization techniques. These authors observed that existing prioritization techniques suffer a lot of setback because they can only handle software projects with very few requirements specifications. Consequently, this has rendered current techniques unsuitable for the prioritization of large scale sets of requirements during software development. (2) the work of Achimugu *et al.* [25] provided another in-depth review that was based on the classification of existing prioritization techniques, and focused on their various setbacks and processes. The authors made some pertinent discoveries and recommendations on the grey areas for enhancement. (3) the work of Pergher and Rossi [26] provided the justification of evaluating prioritization tools by conducting a methodological mapping study. This was further supported by Dabbagh *et al.* [27] with focus on executing two consecutive controlled experiments that aimed at evaluating current prioritization techniques. (4) the work of Rinkēvičs and Torkar [28] that focused on the empirical analysis of commutative voting and the corresponding outcome. The authors proposed the ECV methods for the analysis of results from the CV technique. Again, this technique suffers some setback in terms of analyzing, negotiating and prioritizing large number of requirements during development.

In most cases, the common result with prioritizing software requirement is an ordering of prioritized lists of requirements that needs to be considered first during the software development process [29]. The justification for the acceptance of software systems by its stakeholders have been based on how well the requirements are captured, analyzed and prioritized [30, 31, 32]. The literature contained a lot of contribution that have been made in requirements prioritization research [33, 34]. This is evident in the renowned number of techniques that have been proposed and implemented at different times, and on different development projects. Common to these techniques is their ability to define requirements with greater value to business successes. Racheva *et al.* [35] and Berander *et al.* [36] provided different categorization of prioritization techniques.

The first categorization by Racheva *et al.* [35] includes two main classes: (1) techniques applicable to small-scale requirements, for example, round-the-group prioritization, multi-voting system, pair-wise analysis, weighted criteria analysis, and the quality function deployment techniques; and (2) techniques useful for large-scale requirements, for example, MoSCoW, binary priority list, planning game, case based rank and the Wiegers's matrix techniques.

The second categorization by Berander *et al.* [36] also provided two main classifications. The first classification was based on techniques that support the assignment of values and/or weights by project stakeholders on each requirement. The essence of this classification is to describe the importance of each requirement comparatively. An example of this

includes the analytical process (AHP), planning game, cumulative voting, numerical assignment, and Wieger's method. The second classification suggests the approaches supporting negotiation. In this case, the priorities of requirements are ascertained from the agreement established among the subjective evaluation given by each stakeholder. An example of this classification is the Win-Win model and multi criteria preference analysis requirement negotiation (MPARN) technique.

However, all the techniques from each categorization given in [35] and [36] have their limitations. For example, with the Analytical Hierarchy Process (AHP) as a technique, an $n \times (n - 1) / 2$ comparison was proposed at each hierarchy level given n requirements. This has been evidently seen as a shortcoming of the AHP. This is because when the number of requirements increases, the number of comparisons will also increase with a magnitude of $O(n^2)$ [37, 38]. In addition, the AHP and CBRanks techniques have demonstrated high capabilities as reported in [5]. Both techniques are easy to use, and their accuracy level is very high. Still, their limitations are based on lack of support for scalability [39, 40, 41], and the inability to support rank reversals. This is obvious when considering a larger number of requirements for developing large scale system like the hospital systems. Detailed work that provided more insight on existing requirements prioritization and their corresponding shortcomings have been reported in [25] and other related works. Therefore, we proposed a more efficient ranking algorithm for requirements prioritization based on the limitations of existing techniques.

III. PROPOSED METHOD

Fig. 1 described the procedures and corresponding processes adapted from Perini et al. [42] for the purpose of prioritizing software

requirements. It consists of two phases which involve the manual and automated phases. The manual phases have to do with the specification and weighting of requirements while the automated process has to do with ranking of the weighted requirements and the display of the final ranked requirement. As shown in fig. 1, the ranking procedure indicating the basic flow of processes involved in prioritizing software requirements consists of some important tasks which must be observed in order to achieve reliable prioritization results. These tasks are enumerated below:

1) Requirement Elicitation: This is the number one task in the process of requirements prioritization. It involves the elicitation, gathering and/or extraction of requirements. To elicit requirements, the elicitor, developer or engineer must articulate the description of problem source and how the proposed software is meant to solve the problem identified to the co-opted stakeholders. These stakeholders must acquire adequate understanding of the problem domain and proposed solution before elicitation commences. In cases where stakeholders come with their own requirements, adequate care must also be taken to analyze each requirement with respect to feasibility of implementation, relevant hardware support availability and compatibility, availability of skilled programmers, budget constraints, delivery time and the overall value of the proposed software to the organization or end-users. To initiate the elicitations process, the stakeholders are led to express their views in short sentences where each stakeholder quietly document requirements. Thereafter, stakeholders engage in a round-robin feedback to collate all the requirements so that a discussion section could be organized to arrive at consensus requirements and resolve ambiguities. The requirement elicitation process is given by Equation 1 below:

$$R_{(s,d)} = \min \sum_{i=1}^d \max \sum_{j=1}^d \alpha R_{(s_i d_j)} \left\{ \begin{array}{l} \alpha = 1; \\ \alpha = n; \end{array} \right. \quad (1)$$

Where:

- i and j are the specified requirements from the first to the last stakeholder, respectively.
- $R_{(s,d)}$ is the total number of elicited requirements from an origin node, “ s ” to a last node, “ d ”.

- α is an integer which is non-negative
- $R_{(s_i d_j)}$ is the total number of stakeholders and their respective specified requirements.

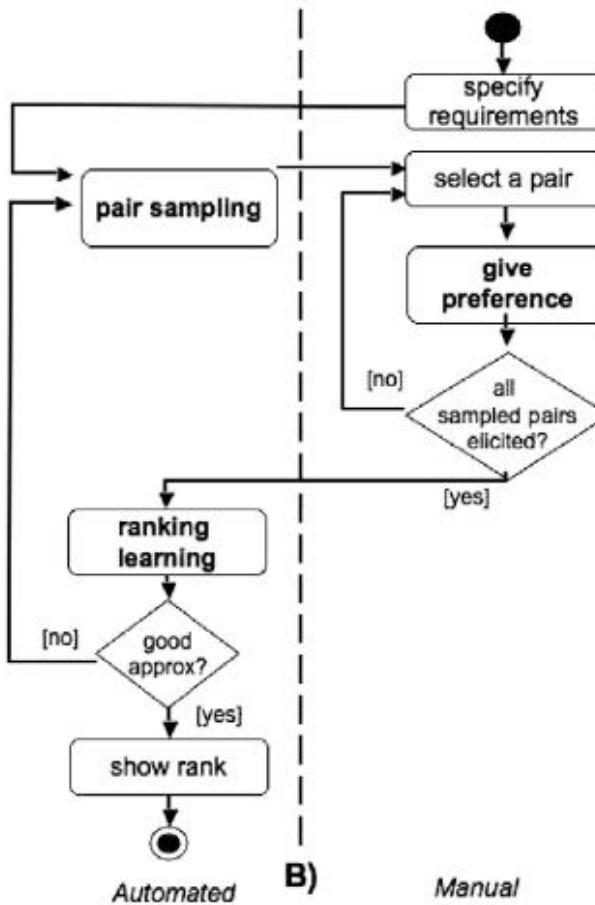


Fig. 1. Model's Information Flow (Adapted from Perini *et al.* [42])

2) Pair Sampling: This task is required in determining the relative importance of a requirement against the other. The relative importance is determined based on pre-defined criteria. Usually, the criteria are set based on the targeted output of the

proposed system. Given some sets of requirements ($r_i, r_j \in R$), the process of pair sampling can be represented with the sets of equations enumerated below:

$$r_{(j-1,j)} \otimes r_{j-1}, r_{(i,j)} = W_j ; \text{ s.t. } i,j \in R \quad (2)$$

$$r_{(j-2,j-1)} \otimes r_{j-2}, r_{(i-1,j-1)} = W_{j-1} ; \text{ s.t. } i,j \in R \quad (3)$$

$$r_{(j-3,j-2)} \otimes r_{j-3}, r_{(i-2,j-2)} = W_{j-2} ; \text{ s.t. } i,j \in R \quad (4)$$

$$\dots \otimes \dots , \dots = \dots ; \text{ s.t. } i,j \in R$$

$$\dots \otimes \dots , \dots = \dots ; \text{ s.t. } i,j \in R$$

$$r_{(1,2)} \otimes W_1, r_{(2,3)} = W_2 ; \text{ s.t. } i,j \in R \quad (5)$$

Where:

- i and j are the consensus requirements from the first to the last
- \otimes is the transitive operator that aid the comparison between one requirement and the other
- r are given sets of requirements
- R is the pool of consensus requirements

3) Preference Elicitation: This is the act of obtaining results of the pair sampling processes from the stakeholders. More precisely, this task is concerned with the collection of all the weighted requirements based on set criteria (c). These weights are eventually used to compute of the final rank of

the entire requirements. The preference elicitation process is executed with the following expressions:

$\otimes : R \times R \rightarrow \{-1, 0, 1\}$ where $\otimes (r_i, r_j)c = 1$ means that r_j has been ranked above r_i ,

$\otimes (r_i, r_j)c = -1$ means that r_i has been ranked above r_j , and

$\otimes (r_i, r_j)c = 0$ indicates that no preference has been given between r_i and r_j

(We assume $\otimes (r_i, r_j)c = 0$ and $\otimes (r_i, r_j)c = -\otimes (r_j, r_i)c$ for all $r_i, r_j \in R$).

4) Ranking Learning: The weights of requirements obtained from (iii) serves as input during ranking learning. The idea here is to compute all the specific ranks of each requirement across all the stakeholders. The learning process relies on the prescribed weights

and ranking criteria to process ranking results. Given some sets of weights; $\mathbf{W}_i, \mathbf{W}_j$, on requirements ($r_i, r_j \in R$); ranking R^* can be executed using the equation 6:

$$R^*(s,d) = \sum_{i=1}^d W_i r_i, W_j r_j \sum_{j=1}^d W_n r_n, W_m r_m R^*(s_i, d_j)$$

s.t. $i & j; n & m \geq 1$; (6)

Where:

- R^* stands for ranked requirements
- s and d are denoted for the start and end of the consensus requirements
- W stands for the weights allotted to each requirement
- r are given sets of requirements
- i and j represents the relative requirements
- n and m represents the end of relative requirements

5) The Final Rank: This is the output of the entire process which reflects the value of each requirement as perceived by the stakeholders. Software requirement specifications deal with the representations of structural components and the relationships that tie them together. In order to avoid vague specification of requirements, we have suggested the application of composition operator (\otimes) that will support pair wise comparison which has the capacity of pre-processing requirements. The synchronization of these representations can be easily achieved if all the components of the proposed

software are well identified and consistently articulated. The preceding description of the proposed requirement prioritization technique forms part of the research contributions and will lead to unambiguous requirements specification.

The inputs to the model consist of the following:

- (a) A set of finite requirements $R = \{r_1, \dots, r_n\}$.
- (b) The ranking criteria $C = (c_1, \dots, c_m)$.
- (c) The stakeholders' preferences represented by the function $\otimes (r_i, r_j) c$

During rank computations, the essence is to generate a priority list of all the requirements contained in R . This priority is represented by the function $P: R \rightarrow R$. The function P stands for the hierarchy of R determined by the preference weights allotted to the various requirements. The aim of any prioritization exercise is to reduce the level of discrepancies or disagreements between prioritized requirements in order to curb ranking error as well as ensure scalable prioritization process. Therefore, if r_i, r_j are given sets of requirements, it is important to ensure the transitive closure $\otimes (r_i, r_j) > 0$. This will ensure non-zero weights on any requirement. The proposed algorithm has the capacity of ordering requirements based on weighted scores.

In fig. 2, we demonstrated that the actual input to the ranking process is a finite set of elicited requirements that are about to be ranked while the ultimate output of the entire process consists of sets of displayed ranked requirements based on pre-defined criteria. The criteria in this context outlined the steps considered suitable in the ranking process as defined in the ranking algorithm.

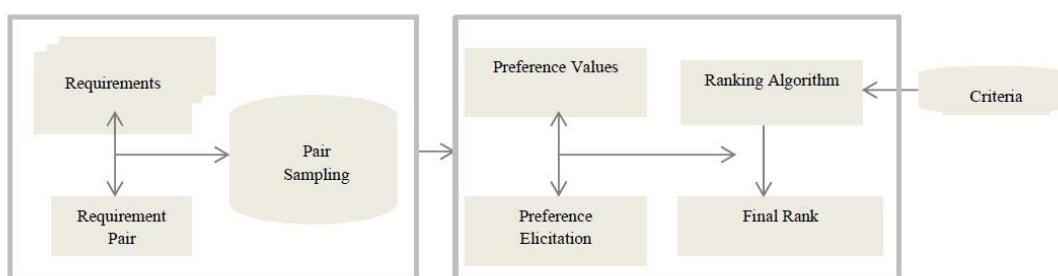


Fig. 2. Ranking-based flow of processes in prioritizing software requirements

In the algorithm, the focus here is to assign membership functions against each requirement based on the allotted weights in order to construct a decision matrix. For instance, assuming we have sets of requirements R elicited from a particular software development project, $R_i (i=1,2,\dots,m)$ which is been evaluated with respect to the number n of selected criteria $C_j (j=1,2,\dots,n)$ that constitute each of those requirements; then, weights can be determined by relevant stakeholders to track the following: (a) the criteria that are to be utilized in ranking the

requirements and (b) construction of the decision matrix using the linguistic variables expressed in Table 1.

TABLE I: Linguistic variables

S/N	Variables	Meaning
1	VL	Very Low
2	L	Low
3	M	Medium
4	H	High
5	VH	Very High

The weighted criteria W stand for the comparative rank of the chosen criteria; while, decision matrix represents the overall ranking of each requirement R_i in line with the prescribed criteria C_j .

Step 1: Construction of the decision matrix X : The weights are used to construct decision matrix of the form shown below:

$$D = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \\ R_1 & \left[\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1n} \end{array} \right] \\ R_2 & \left[\begin{array}{cccc} x_{21} & x_{22} & \cdots & x_{2n} \end{array} \right] \\ \vdots & \vdots & \ddots & \vdots \\ R_m & \left[\begin{array}{cccc} x_{m1} & x_{m2} & \cdots & x_{mn} \end{array} \right] \end{bmatrix} \quad (7)$$

$$W = [w_1 \quad w_2 \quad \cdots \quad w_n]$$

After constructing decision matrix, there is need to normalize the matrix by using Equation 8.

$$R_{ij} = \sum_{j=1}^i x_{ij} \quad i = 1, \dots, n; j = 1, \dots, m \quad (8)$$

Step 2: Aggregation of normalized decision weights: The aggregated weights for the normalized entries are computed by obtaining the square root of each of the normalized weight using Equation 9.

$$W_j = \sqrt[m]{w_{1,j} \cdots w_{m,j}}, \quad j = 1, \dots, n \quad (9)$$

Step 3: Computation of the aggregated fuzzy decision matrix in its linguistic form using Equation 10.

$$R_{ij} = \sum_{j=1}^i x_{ij}^2 \quad i = 1, \dots, n; j = 1, \dots, m \quad (10)$$

Step 4: Computation of the global weights: The i th values in Step 3 are multiplied by each values of the aggregated decision matrix in Step 2 using Equation 11 to acquire the global weights of each criterion.

$$W_j = \sqrt[m]{w_{1,j} \times w_{2,j} \times w_{3,j} \cdots w_{m,j}} \times \sum_{j=1}^i x_{ij}^2, \quad i = 1, \dots, n; j = 1, \dots, m \quad (11)$$

Step 5: Computation of the Positive Ideal Solution (PIS) and Negative Ideal Solution (NIS): The solutions from the global decision matrix values in Step 4 are used to compute the PIS and NIS. To compute the PIS, Equation 12 is utilized.

$$A^* = (v_1^*, v_2^*, \dots, v_n^*) \quad (12)$$

Where, v_j^* = the maximum values of the requirement entries in the aggregated decision matrix. NIS is computed with Equation 13.

$$A' = (v'_1, v'_2, \dots, v'_n) \quad (13)$$

Where, v' = the minimum values of the requirement entries in the aggregated decision matrix.

Step 6: Computation of the separating distances: This is achieved by finding the variance amidst the highest and lowest values contained in the aggregated decision matrix as depicted in Equation 14.

$$S_i = \left(\sum (A^*_{ij} - A^-_{ij}) \right) \quad (14)$$

Where $i = 1, 2, \dots, m$

Step 7: The results in step 6 represent the confidence rating of each requirement with respect to the relevant stakeholders. Therefore, the requirement with the highest confidence rating (CR) value is considered as the prime requirement.

In dealing with decision making challenges during software elicitation process, it is necessary to consider the number of requirements specified by stakeholders (or the number of criteria that makes up each requirement) in order to find the Euclidean distances amongst positively ideal requirements. Assuming a particular software engineer have acquired two major requirements for a certain software development project, say (R_1, R_2) ; then it becomes lighter to determine the PIS criteria which consist of all highly ranked criteria that constitute a requirement. Conversely, the NIS which consists of criteria that are least ranked is also detected. However, if two requirements R_1 and R_2 possess a shorter distance to both PIS and NIS; it becomes quite imperative to clearly define the rationale for choosing one over the other or choosing both as the case may be. This algorithm deals with the concept of considering alternatives, known as compromise solution, that possesses the nearest distance to the PIS and the farthest distance from the NIS.

IV. EXPERIMENTAL SET-UP

To illustrate the concept of these techniques, the Pharmacy Information System at the Obafemi Awolowo University Teaching Hospital Complex (OAUTHC) was considered in this case. Consider the following scenario: "The Pharmacy Department at OAUTHC would like to develop new software to replace the existing one so that a better solution for the pharmacy operation is achieved, and each pharmacy sub-unit's functions and activities are captured. The new system should facilitate the administration of both outpatient and inpatient medication supplies to the wards, and also provides inventory support to manage stock movement from any given location across the three tiers of healthcare delivery system in Nigeria".

The system must be flexible enough to enable Pharmacists and other experts gain access into the system and administer the appropriate healthcare services. Bearing in mind that, OAUTHC and other hospitals deal with complex and large amount of data; there would be a need to build a scalable system that will be dynamic and interoperable. Thus, the need to provide an appropriate measure of ranking requirements in a manner to avoid delay in implementation, reduce cost of development, and ensure quality assurance and control. The system should also be able to assist in patient care by the monitoring of drug interactions, drug allergies and other possible medication-related complications, in addition to capturing relevant medical information.

The essence of capturing this medical information of patients could be aimed at indexing into a well-structured database that is void of redundancy or replications. Basically, the software system can be developed to enable Pharmacist to manage medication orders, ensure that preparations, dispensing, and verification are all easily completed and monitored.

From the above scenario, the software engineers and other relevant stakeholders are trying to gather information in order to ensure that, the proposed system works as expected. Nine stakeholders are involved in this case. Based on the elicited information, the architectural design decisions are made. Due to several reasons, architectural decision-making is a difficult task because: (i) requirements are usually captured with a lot of ambiguous insinuations (ii) functional and non-functional requirements are difficult to explicitly specify (iii) concrete architectural decisions have to be made based on vague requirements in some cases (iv) stakeholders involved have different perspectives, views or concerns.

Possessing a mastery of sound elicitation technique and being able to clearly describe problems leads to concise specification of requirements. This is the key for developing a formidable architecture for software systems. To get started, software engineers and stakeholders must be able to choose views, sort them, prioritize and document them which translate to the following sequence of activities:

- a. The utilization of an elicitation technique must lead to unambiguous requirements specification.
- b. The architectural design must conform to the elicited requirements.
- c. The quality attributes of the proposed system are selected from the non-functional requirements class.
- d. The mapping between architectural design decisions and requirement specifications must be consistent.

During or after requirements specification, it is important to collapse them into categories as described in the hierarchical representation of the specified requirements shown in fig 3. This will enhance clarity in the evaluation process; since these requirements are about to be prioritized. When prioritization is to commence, requirements are compared pair-wisely where relevant stakeholders are required to determine the relative importance of each requirement based on the comparison scale. The requirements undergoing pair wise comparison are performance, reusability, flexibility and maintainability; denoted as R_1, R_2, R_3 and R_4 respectively.

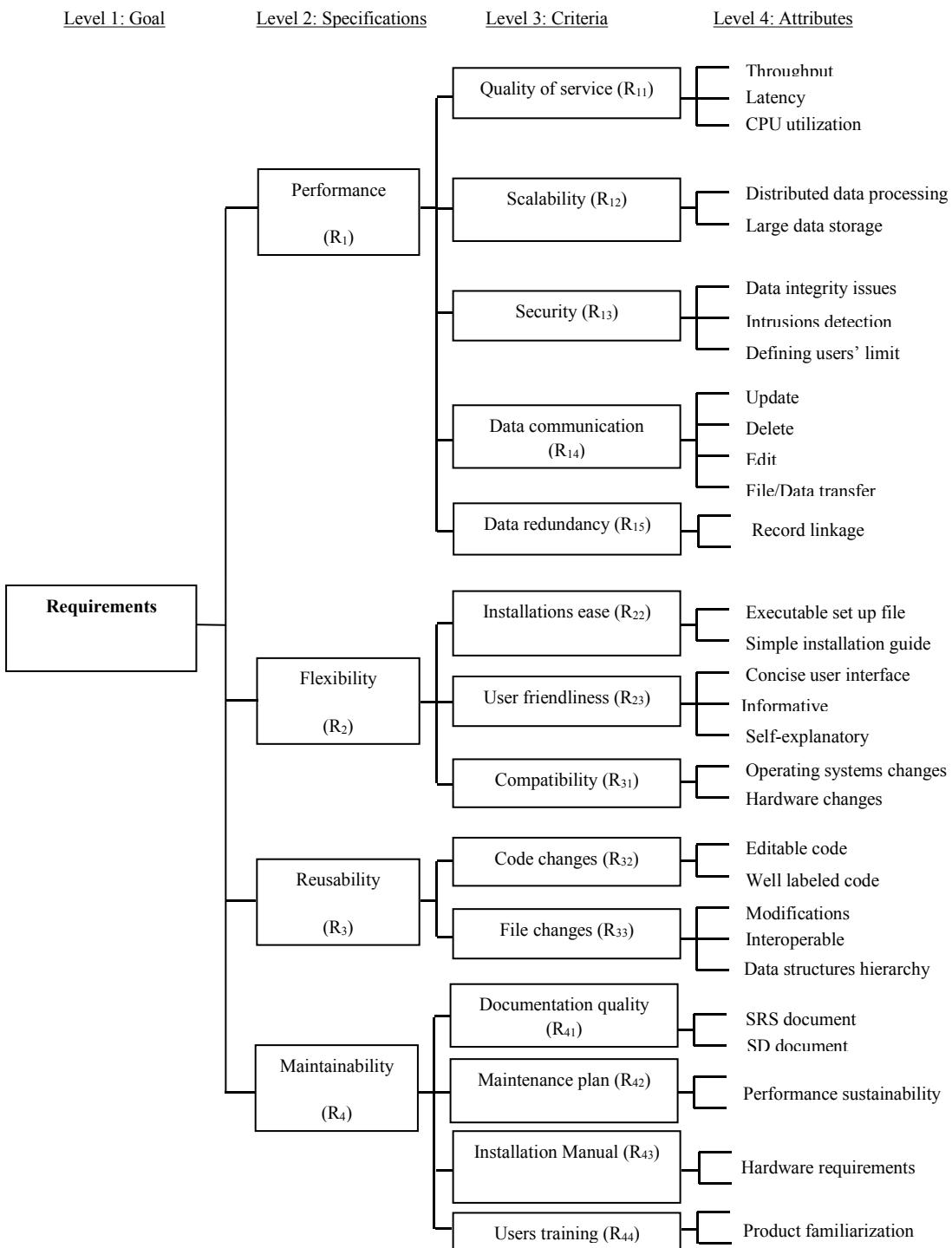


Fig. 3. Categorization of Stakeholder's requirements

In the context of the Pharmacy information system from literature, performance has to do with the overall deliveries supported by a system. However, the attributes that constitute this requirement include quality of service (system availability); scalability (large amount of data processing at runtime) and Security (protection of patient data against malicious or unauthorized users), data communication (basic database updates across distributed database applications) and data redundancy (ability of a distributed database system to detect or avoid data replications by linking records to appropriate entities).

The requirement tagged flexibility on the other hand has to do with the level of portability of the system. That is, the ability of the system to be installed and used across any operating system platform. Precisely, the attributes under this category are installation ease (ability for the system to be easily installed on independent platforms); user friendliness (a system with concise and unambiguous features or functions). Interfaces are expected to be designed with clear links or Webpages and compatibility; that is, the ability of the system to adapt to sudden changes in technological or infrastructural advancement.

Reusability measures the extent at which a system can be long lived; that is, the degree at which a system can evolve or adapt to operating system changes or modifications, as well as data, algorithm and file changes. This is crucial because, a system that is not reusable will not have any future value. If a system ends up not having any future value, it creates an impression that money has been wasted in the project. Moreover, organisations in recent times are beginning to appreciate systems that possess the ability to adapt to future changes with little or no administrative interventions.

Finally, the maintenance requirement is responsible for prolonging or sustaining the life span of the developed software systems. This involves the provision of the software requirement and design documents as well as the application codes, a periodic maintenance plan and a user training session to enable users get themselves acquainted with the software system and installation manual which will help guide users to a successful installation of the software. This could also help them perform simple maintenance exercises.

V. RESULTS AND DISCUSSION

Table 2 indicates how the subjective priorities given to each requirement by the stakeholders were captured. To achieve that, the linguistic valuations were introduced. In this case, Ranking Scales (RS) were used as variables. The RS are assertions that use expressions in English language to represent values that stands for the degree of acceptability of a particular variable or parameter. One advantage of the linguistic variables is the ability to determine the level of acceptability amongst specified requirements by each stakeholder. Therefore, using the RS, stakeholders were able to provide preference weights for the elicited requirements based on their perceived importance.

Next to the priority list indicated by the RS is Tables 3 and 4 showing the ranking with normalised and aggregate weights derived from equation 8 and 9 respectively. These normalised decision weights determine the performance of ranking algorithm. Tables 5 and 6 showed the Fuzzy and Global decision matrices derived from equation 10 and 11 respectively. The fuzzy logic concept was employed in [1] as a way of overcoming complexities in decision making. Consequently, equation 12 was used to compute the positive ideal solution (PIS) and negative ideal solution (NIS), as well as the closeness coefficient (CC).

The two ideal solutions (PIS and NIS) were computed for 4 requirements across 3 stakeholders as shown in Tables 7 and 8 respectively. PIS and NIS was introduced to provide intuitive and computationally feasible approach to identify the CC. Hence, the CC showing the final ranks of requirements are depicted in Table 9. The CC

provides the medium through which problems of multiple criteria decision making can be handled, so as to determine the order of priority of available alternatives. Experimentally, the CC showed that R4 with 2.09 point is the most valued requirements. Next to R4 in the order of priority is R1 and R2 respectively with CC of 1.37 and 1.05 respectively. Thus, we are optimistic that this ranking algorithm can cater for large dataset (i.e. large number of requirements). As such, prioritizing software requirements with the suggested ranking algorithm during software development can reduce development cost, and aid smooth release management of software products timely.

Therefore, the results in Tables 2 to 9 represent the performance of the proposed technique. These results emphasized more on scalability and ranking evolving requirements. The advantages of the ranking algorithm are as follows:

- (a) Scalability: The algorithm can be applied to requirements of ultra-large scale software development projects since the weights of requirements can be computed across all their respective criteria at runtime.
- (b) Evolvability (Rank reversal): The algorithm also has the capacity of accurately calculating weights across even or odd numbers of criteria that constitute each requirement. The functions for computing PIS and NIS have the capacity of computing relative weights between odd or even requirements.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a more efficient ranking algorithm that is capable of resolving redundant specified requirements. The ranking algorithm proposed, demonstrated the potential of catering for large number of requirements as datasets. Secondly, it has the potential of facilitating effective and efficient decision making that can handle the differences amongst requirements that have been prioritized. Therefore, it is believed that this approach can help software engineers to prioritize requirements capable of forecasting the expected behaviour of software under development. The implication of the final ranked requirements would be that those that emerged top in the ranked list will be implemented first while others can be implemented subsequently in the software release planning process.

The next phase of this research can be tailored towards the minimization of discrepancies between ranked requirements. Secondly, the automation of the computations proposed in this paper and the display of ranked results will be required. Finally, it will be so interesting to validate this proposed technique with a real life software development project and applicable datasets as well as implementation of the prototype.

TABLE II: Linguistic variables

R₁	R ₁₁	R ₁₂	R ₁₃	R ₁₄	R ₁₅	R₂	R ₂₁	R ₂₂	R ₂₃	R₃	R ₃₁	R ₃₂	R₄	R ₄₁	R ₄₂	R ₄₃	R ₄₄
S ₁	VH	VH	VH	VH	VH	S ₁	H	H	H	S ₁	H	H	S ₁	VH	VH	VH	VH
S ₂	VH	VH	VH	VH	VH	S ₂	VH	VH	VH	S ₂	VH	VH	S ₂	VH	VH	VH	VH
S ₃	H	H	H	H	H	S ₃	VH	VH	VH	S ₃	VH	VH	S ₃	M	M	M	M
S ₄	H	H	H	H	H	S ₄	M	M	M	S ₄	VH	VH	S ₄	VH	VH	VH	VH
S ₅	VH	VH	VH	VH	VH	S ₅	H	H	H	S ₅	M	M	S ₅	H	H	H	H
S ₆	H	H	H	H	H	S ₆	VH	VH	VH	S ₆	M	M	S ₆	H	H	H	H
S ₇	VH	VH	VH	VH	VH	S ₇	VH	VH	VH	S ₇	M	M	S ₇	H	H	H	H
S ₈	H	H	H	H	H	S ₈	M	M	M	S ₈	H	H	S ₈	H	H	H	H
S ₉	VH	VH	VH	VH	VH	S ₉	H	H	H	S ₉	M	M	S ₉	VH	VH	VH	VH

TABLE III: Normalized weights

R₁	R ₁₁	R ₁₂	R ₁₃	R ₁₄	R ₁₅	R₂	R ₂₁	R ₂₂	R ₂₃	R₃	R ₃₁	R ₃₂	R₄	R ₄₁	R ₄₂	R ₄₃	R ₄₄
S ₁	0.10	0.10	0.10	0.10	0.10	S ₁	0.06	0.06	0.06	S ₁	0.06	0.06	S ₁	0.10	0.10	0.10	0.10
S ₂	0.10	0.10	0.10	0.10	0.10	S ₂	0.10	0.10	0.10	S ₂	0.10	0.10	S ₂	0.10	0.10	0.10	0.10
S ₃	0.06	0.06	0.06	0.06	0.06	S ₃	0.10	0.10	0.10	S ₃	0.10	0.10	S ₃	0.02	0.02	0.02	0.02
S ₄	0.06	0.06	0.06	0.06	0.06	S ₄	0.02	0.02	0.02	S ₄	0.10	0.10	S ₄	0.10	0.10	0.10	0.10
S ₅	0.10	0.10	0.10	0.10	0.10	S ₅	0.06	0.06	0.06	S ₅	0.02	0.02	S ₅	0.06	0.06	0.06	0.06
S ₆	0.06	0.06	0.06	0.06	0.06	S ₆	0.10	0.10	0.10	S ₆	0.02	0.02	S ₆	0.06	0.06	0.06	0.06
S ₇	0.10	0.10	0.10	0.10	0.10	S ₇	0.10	0.10	0.10	S ₇	0.02	0.02	S ₇	0.06	0.06	0.06	0.06
S ₈	0.06	0.06	0.06	0.06	0.06	S ₈	0.02	0.02	0.02	S ₈	0.06	0.06	S ₈	0.06	0.06	0.06	0.06
S ₉	0.10	0.10	0.10	0.10	0.10	S ₉	0.06	0.06	0.06	S ₉	0.02	0.02	S ₉	0.10	0.10	0.10	0.10

TABLE IV: Aggregate weights

R₁	R ₁₁	R ₁₂	R ₁₃	R ₁₄	R ₁₅	R₂	R ₂₁	R ₂₂	R ₂₃	R₃	R ₃₁	R ₃₂	R₄	R ₄₁	R ₄₂	R ₄₃	R ₄₄
S ₁	0.316	0.316	0.316	0.316	0.316	S ₁	0.245	0.245	0.245	S ₁	0.245	0.245	S ₁	0.316	0.316	0.316	0.316
S ₂	0.316	0.316	0.316	0.316	0.316	S ₂	0.316	0.316	0.316	S ₂	0.316	0.316	S ₂	0.316	0.316	0.316	0.316
S ₃	0.245	0.245	0.245	0.245	0.245	S ₃	0.316	0.316	0.316	S ₃	0.316	0.316	S ₃	0.141	0.141	0.141	0.141
S ₄	0.245	0.245	0.245	0.245	0.245	S ₄	0.141	0.141	0.141	S ₄	0.316	0.316	S ₄	0.316	0.316	0.316	0.316
S ₅	0.316	0.316	0.316	0.316	0.316	S ₅	0.245	0.245	0.245	S ₅	0.245	0.245	S ₅	0.245	0.245	0.245	0.245
S ₆	0.245	0.245	0.245	0.245	0.245	S ₆	0.316	0.316	0.316	S ₆	0.141	0.141	S ₆	0.245	0.245	0.245	0.245
S ₇	0.316	0.316	0.316	0.316	0.316	S ₇	0.316	0.316	0.316	S ₇	0.141	0.141	S ₇	0.245	0.245	0.245	0.245
S ₈	0.245	0.245	0.245	0.245	0.245	S ₈	0.141	0.141	0.141	S ₈	0.245	0.245	S ₈	0.245	0.245	0.245	0.245
S ₉	0.316	0.316	0.316	0.316	0.316	S ₉	0.245	0.245	0.245	S ₉	0.141	0.141	S ₉	0.316	0.316	0.316	0.316

TABLE V: Fuzzy decision matrix

R₁				R₂				R₃				R₄			
S ₁	0.80	1.25	1.25	S ₁	0.27	0.48	0.75	S ₁	0.18	0.32	0.50	S ₁	0.64	1.00	1.00
S ₂	0.80	1.25	1.25	S ₂	0.48	0.75	0.75	S ₂	0.32	0.50	0.50	S ₂	0.64	1.00	1.00
S ₃	0.45	0.80	1.25	S ₃	0.48	0.75	0.75	S ₃	0.32	0.50	0.50	S ₃	0.16	0.36	0.64
S ₄	0.45	0.80	1.25	S ₄	0.12	0.27	0.48	S ₄	0.32	0.50	0.50	S ₄	0.64	1.00	1.00
S ₅	0.80	1.25	1.25	S ₅	0.27	0.48	0.75	S ₅	0.08	0.18	0.32	S ₅	0.36	0.64	1.00
S ₆	0.45	0.80	1.25	S ₆	0.48	0.75	0.75	S ₆	0.08	0.18	0.32	S ₆	0.36	0.64	1.00
S ₇	0.80	1.25	1.25	S ₇	0.48	0.75	0.75	S ₇	0.08	0.18	0.32	S ₇	0.36	0.64	1.00
S ₈	0.45	0.80	1.25	S ₈	0.12	0.27	0.48	S ₈	0.18	0.32	0.50	S ₈	0.36	0.64	1.00
S ₉	0.80	1.25	1.25	S ₉	0.27	0.48	0.75	S ₉	0.08	0.18	0.32	S ₉	0.64	1.00	1.00

TABLE VI: Global fuzzy decision matrix

R₁				R₂				R₃				R₄			
S ₁	1.00	1.57	1.57	S ₁	0.26	0.47	0.73	S ₁	0.13	0.22	0.35	S ₁	0.72	1.12	1.12
S ₂	1.00	1.57	1.57	S ₂	0.47	0.73	0.73	S ₂	0.25	0.40	0.40	S ₂	0.72	1.12	1.12
S ₃	0.50	0.89	1.38	S ₃	0.47	0.73	0.73	S ₃	0.25	0.40	0.40	S ₃	0.12	0.27	0.48
S ₄	0.50	0.89	1.38	S ₄	0.15	0.26	0.47	S ₄	0.25	0.40	0.40	S ₄	0.72	1.12	1.12
S ₅	1.00	1.57	1.57	S ₅	0.26	0.47	0.73	S ₅	0.04	0.10	0.17	S ₅	0.36	0.63	0.99
S ₆	0.50	0.89	1.38	S ₆	0.47	0.73	0.73	S ₆	0.04	0.10	0.17	S ₆	0.36	0.63	0.99
S ₇	1.00	1.57	1.57	S ₇	0.47	0.73	0.73	S ₇	0.04	0.10	0.17	S ₇	0.36	0.63	0.99
S ₈	0.50	0.89	1.38	S ₈	0.15	0.26	0.47	S ₈	0.13	0.22	0.35	S ₈	0.36	0.63	0.99
S ₉	1.00	1.57	1.57	S ₉	0.26	0.47	0.73	S ₉	0.04	0.10	0.17	S ₉	0.72	1.12	1.12

TABLE VII: Positive ideal solution

				d^+
R ₁	1.00	1.57	1.57	4.14
R ₂	0.47	0.73	0.73	1.93
R ₃	0.25	0.40	0.40	1.05
R ₄	0.72	1.12	1.12	2.96

TABLE VIII: Negative ideal solution

				d^-
R ₁	0.50	0.89	1.38	2.77
R ₂	0.15	0.26	0.47	0.88
R ₃	0.04	0.10	0.17	0.31
R ₄	0.12	0.27	0.48	0.87

TABLE IX: Closeness coefficient

	d^+	d^-	CC	Ranking
R ₁	4.14	2.77	1.37	2
R ₂	1.93	0.88	1.05	3
R ₃	1.05	0.31	0.74	4
R ₄	2.96	0.87	2.09	1

ACKNOWLEDGEMENTS

The Health Informatics Research Group (HIRG) of Computer Science and Engineering, Faculty of Technology, Obafemi Awolowo University, Ile-Ife, Nigeria is hereby acknowledged for some of the facilities utilized during the course of this research.

REFERENCE

- [1] I. Gambo, R. Ikono, P. Achimugu, and A. Soriyan, "An Integrated Framework for Prioritizing Software Specifications in Requirements Engineering," *International Journal of Software Engineering and its Applications (IJSEIA)*, Vol. 12, No. 1, 2018, pp. 33-46. ISSN 1738-9984.
- [2] P. Achimugu, and A. Selamat, "A Hybridized Approach for Prioritizing Software Requirements Based on K-Means and Evolutionary Algorithms," *Computational Intelligence Applications in Modelling and Control*, 2015, pp. 73-93. Springer International Publishing.
- [3] P. Achimugu, A. Selamat, and R. Ibrahim, "ReProTizer: A Fully Implemented Software Requirements Prioritization Tool," *Transactions on Computational Collective Intelligence*, XXII, 2016, pp. 80-105. Springer Berlin Heidelberg.
- [4] S. Liaskos, S. A. McIlraith, S. Sohrabi, and J. Mylopoulos, "Representing and reasoning about preferences in requirements engineering," *Requirements Engineering*, Vol. 16, No. 3, 2011, pp. 227-249.
- [5] M. I. Babar, M. Ghazali, D. N. Jawawi, S. M. Shamsuddin, and N. Ibrahim, "PHandler: An expert system for a scalable software requirements prioritization process," *Knowledge-Based Systems*, Vol. 84, 2015, pp. 179-202.
- [6] G. Ruhe, A. Eberlein, and D. Pfahl, "Trade-off Analysis for Requirements Selection," *International Journal of Software Engineering and Knowledge Engineering*, Vol. 13, No. 4, 2003, pp. 345-366.
- [7] R. K. Chopra, V. Gupta, and D. S. Chauhan, "Experimentation on Accuracy of Non Functional Requirement Prioritization Approaches for different complexity Projects," *Perspectives in Science*, 2016, <http://dx.doi.org/10.1016/j.pisc.2016.04.001>.
- [8] A. Finkelstein, M. Harman, S. Mansouri, J. Ren, and Y. Zhang, "A search based approach to fairness analysis in requirement assignments to aid negotiation, mediation and decision making," *Requirements Engineering*, Vol. 14, 2009, pp. 231–245.
- [9] S. Barney, A. Aurum, and C. Wohlin, "A product management challenge: creating software product value through requirements selection," *Journal of System Architecture*, Vol. 54, 2008, pp. 576–593.
- [10] P. Belsis, A. Koutoumanos, and C. Sgouropoulou, "PBURC: a patterns-based, unsupervised requirements clustering framework for distributed agile software development," *Requirements Engineering*, Vol. 19, No. 2, 2014, pp. 213-225.
- [11] A. S. Jadhav, and R. M. Sonar, "Framework for evaluation and selection of the software packages: A hybrid knowledge based system approach," *Journal of Systems and Software*, Vol. 84, No. 8, 2011, pp. 1394-1407.

- [12] L. Pareto, A. B. Sandberg, P. Eriksson, and S. Ehnebom, "Collaborative prioritization of architectural concerns," *Journal of Systems and Software*, Vol. 85, No. 9, 2012, pp. 1971-1994.
- [13] P. Tonella, A. Susi, and F. Palma, "Interactive requirements prioritization using a genetic algorithm," *Information and software technology*, Vol. 55, No. 1, 2013, pp. 173-187.
- [14] J. Karlsson, S. Olsson, and K. Ryan, "Improved practical support for large-scale requirements prioritizing," *Requirements Engineering*, Vol. 2, No. 1, 1997, pp. 51-60.
- [15] A. De Lucia, and A. Qusef, "Requirements engineering in agile software development," *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, 2010, pp. 212-220.
- [16] B. Ramesh, L. Cao, and R. Baskerville, "Agile requirements engineering practices and challenges: an empirical study," *Information Systems Journal*, Vol. 20, No. 5, 2010, pp. 449-480.
- [17] M. Daneva, E. Van Der Veen, C. Amrit, S. Ghaisas, K. Sikkel, R. Kumar, and R. Wieringa, "Agile requirements prioritization in large-scale outsourced system projects: An empirical study," *Journal of systems and software*, Vol. 86, No. 5, 2013, pp. 1333-1353.
- [18] Z. Bakalova, M. Daneva, A. Herrmann, and R. Wieringa, "Agile requirements prioritization: What happens in practice and what is described in literature," In *Proceedings of the 17th International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ 2011)*, 28th - 30th March, Essen, Germany, 2011, pp. 181-195. Springer, Berlin, Heidelberg.
- [19] I. Inayat, S.S. Salim, S. Marczak, M. Daneva, and S. Shamshirband, "A systematic literature review on agile requirements engineering practices and challenges," *Computers in human behavior*, Vol. 51, 2015, pp. 915-929.
- [20] S. Vinay, S. Aithal, and G. Sudhakara, "A quantitative approach using goal-oriented requirements engineering methodology and analytic hierarchy process in selecting the best alternative," In *Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 22 Aug - 25 Aug 2013, Mysore, India, 2013, pp. 441-454.
- [21] M. Sadiq, and S. K. Jain, "Stakeholder identification method in goal oriented requirements elicitation process," In *proceedings of the IEEE 5th International Workshop on Requirements Prioritization and Communication (RePriCo, 2014)*, 26-26 August, Karlskrona, Sweden, 2014, pp. 25-33.
- [22] M. Sadiq, and S. K. Jain, "A fuzzy based approach for the selection of goals in goal oriented requirements elicitation process," *International Journal of System Assurance Engineering and Management*, Vol. 6, No. 2, 2015, pp. 157-164.
- [23] M. Sadiq, T. Hassan, and S. Nazneen, AHP_GORE_PSR: applying analytic hierarchy process in goal oriented requirements elicitation method for the prioritization of software requirements. In *the Proceedings of the IEEE 3rd International conference on Computational Intelligence & Communication Technology (CICT, 2017)*, 9th - 10th February, Ghaziabad, India, 2017, pp. 1-5.
- [24] A. M. Pitangueira, R. S. P. Maciel, and M. Barros, "Software requirements selection and prioritization using SBSE approaches: A systematic review and mapping of the literature," *Journal of Systems and Software*, Vol. 103, 2015, pp. 267-280.
- [25] P. Achimugu, A. Selamat, R. Ibrahim, and M. N. R. Mahrin, "A systematic literature review of software requirements prioritization research," *Information and Software Technology*, Vol. 56, 6, 2014, pp. 568-585.
- [26] M. Pergher, and B. Rossi, "Requirements prioritization in software engineering: a systematic mapping study," In: *Paper presented at the 2013 IEEE Third International Workshop on Empirical Requirements Engineering (EmpiRE)*, 2013.
- [27] M. Dabbagh, S. P. Lee, and R. M. Parizi, "Functional and non-functional requirements prioritization: empirical evaluation of IPA, AHP-based, and HAM-based approaches," *Soft Computing*, 2015, pp. 1-24.
- [28] K. Rinkēvičs, and R. Torkar, "Equality in cumulative voting: A systematic review with an improvement proposal," *Information and Software Technology*, Vol. 55, No. 2, 2013, pp. 267-287.
- [29] M. Dabbagh, and S. P. Lee, "An approach for prioritizing NFRs according to their relationship with FRs," *Lecture Notes on Software Engineering*, Vol. 3, No. 1, 2015, pp. 1-5.
- [30] A. M. Davis, "The art of requirements triage", *IEEE Computer*, Vol. 36, No. 3, 2003, pp. 42-49.
- [31] N. R. Mead, "Requirements Prioritization Introduction," *Software Engineering Institute Web Publication, Carnegie Mellon University, Pittsburgh, USA*, 2006.
- [32] A. Perini, F. Ricca, A. Susi, and C. Bazzanella, "An empirical study to compare the accuracy of AHP and CB Ranking techniques for requirements prioritization," in: *Proceedings of the Fifth International Workshop on Comparative Evaluation in Requirements Engineering, IEEE*, 2007, pp. 23-35.
- [33] T. L. Saaty, *The analytic hierarchy process*, McGraw-Hill, New York, 1980.
- [34] A. Herrmann, and M. Daneva, "Requirements prioritization based on benefit and cost prediction: an agenda for future research," In: *RE, IEEE Computer Society*, 2008, pp. 125-134.

- [35] Z. Racheva, M. Daneva, A. Herrmann, and R. J. Wieringa, "A conceptual model and process for client-driven agile requirements prioritization," In: *IEEE (2010) Fourth International Conference on Research Challenges in Information Science (RCIS)*, 2010, pp. 287–298.
- [36] P. Berander, K. A. Khan, and L. Lehtola, "Towards a research framework on requirements prioritization," *SERPS 6*, 2006, pp. 18–19.
- [37] J. R. Hubbard, "*Theory and Problems of Data Structures with C++*". McGraw-Hill, NY, USA, 2000.
- [38] N. W. Kassel, and B. A. Malloy, "An approach to automate requirements elicitation and specification," In *the Proceedings of 7th International Conference on Software Engineering and Applications*, 2003, pp. 3-5.
- [39] J. Karlsson, and K. Ryan, "Cost-value approach for prioritizing requirements" *IEEE Software*, Vol. 14, No. 5, 1997, pp. 67-74.
- [40] J. Karlsson, C. Wohlin, and B. Regnell, "An evaluation of methods for prioritizing software requirements," *Information and Software Technology*, Vol. 39, No. 14-15, 1998, pp. 939-947.
- [41] C. Duan, P. Laurent, J. Cleland-Huang, and C. Kwiatkowski, "Towards automated requirements prioritization and triage, Requirements Engineering," Vol. 14, No. 2, 2009, pp. 73–89.
- [42] A. Perini, F. Ricca, and A. Susi, "Tool-supported requirements prioritization: Comparing the AHP and CBRank methods," *Information and Software Technology*, Vol. 51, No. 6, 2009, pp. 1021-1032.

Segmentation of Diffusion Tensor Brain Tumor Images using Fuzzy C-Means Clustering

Ceena Mathews

Department of Computer Science, Prajyoti Niketan College

Pudukad, Thrissur, Kerala, India

ceenamathews@gmail.com

Abstract: A malignant tumor, also called brain cancer, grows rapidly and often invades or crowds healthy areas of the brain. Brain tumors can affect white matter fibers by either infiltrating or displacing the tissue. When the myelin sheath is damaged or disappears, the conduction of impulses along nerve fibers slows down or fails completely. Diffusion Tensor Imaging (DTI) is a relatively new imaging technique that can be used to evaluate white matter in the brain. DTI has diagnostic implications by being able to pinpoint areas where normal water flow is disrupted, providing valuable information about the location of specific lesions. Edema, infiltration and destruction of white matter reduces the anisotropic nature of the white matter. The paper aims to segment tumor from the healthy brain tissues in Diffusion Tensor brain tumor images using Fuzzy C-Means clustering

Tumor; diffusion tensor image; edema; anisotropy; clustering (key words)

I. INTRODUCTION

A brain tumor is a group of abnormal cells that grows in or around the brain. Tumors can directly destroy healthy brain cells. According to WHO, there are an estimated 240,000 cases of brain and nervous system tumors per year, worldwide.

Brain tumors are either malignant or benign. A malignant tumor, also called brain cancer, grows rapidly and often invades or crowds healthy areas of the brain. Benign brain tumors do not contain cancer cells and are usually slow growing. Brain tumors fall into two different categories: primary or metastatic. Primary brain tumors begin within the brain. A metastatic tumor is formed when cancer cells located elsewhere in the body break away and travel to the brain.

The brain of human consists of gray matter and white matter. The gray matter contains the nerve cells. The white matter of the brain is composed of nerve fibers and myelin. The nerve fibers form the connections between the nerve cells. Myelin is an essential part of the white matter. Brain tumors can affect white matter fibers by either infiltrating or displacing the tissue. When the myelin sheath is damaged or disappears, the conduction of impulses along nerve fibers slows down or fails

completely. Consequently, brain functions become hampered or be lost.

Neuroimaging techniques are used to produce images of the brain. Each technique conveys distinct types of information depending on the question at hand. MRI is the primary imaging modality in brain tumor patients. Diffusion Tensor Imaging (DTI) is a relatively new imaging technique that can be used to evaluate white matter in the brain. DTI has diagnostic implications by being able to pinpoint areas where normal water flow is disrupted, providing valuable information about the location of specific lesions.

Brain tumors alter regional brain architecture due to differences in cell structure, size, and density and the presence of necrosis and edema. Consequently, tumor MR diffusion properties may identify diagnostic intertumoral differences. Whole-brain maps of diffusion metrics can be generated from diffusion tensor imaging (DTI) data. Mean diffusivity (MD) provides a magnitude of isotropic diffusion (in $\text{mm}^2 \text{s}^{-1}$), and fractional anisotropy (FA) provides a scalar value of diffusion directionality. Differences in MD and FA among tumor types and grades of malignancy have been investigated with mixed success.

The diffusion of water within the tissues will be altered by changes in the tissue microstructure and organization. Edema, infiltration and destruction of white matter reduces the anisotropic nature of the white matter. DTI can delineate gross abnormality in the white matter anatomy better than conventional MRI. DTI can also be used to differentiate tumor-infiltrated edema from pure vasogenic edema, which may be beneficial for accurate preoperative diagnosis of glioblastomas and metastasis. DTI can be used to differentiate between recurrent tumor and radiation necrosis as in [11].

II. RELATED STUDIES

Diffusion tensor imaging (DTI) has become one of the most popular MRI techniques in brain research, as well as in clinical practice.

A significant number of studies have attempted to use DTI to more precisely delineate the margins of brain

tumors in humans and detect changes in the normal-appearing tissue surrounding malignant gliomas that are not detectable on conventional MR imaging. DTI has been utilized for evaluating the peritumoral region of brain tumors is discussed in [1].

Conventional structural MR modalities are combined with diffusion tensor imaging data to create an integrated multimodality profile for brain tumors. This framework is discussed in [2]. DTI-based histogram and fDM analysed for evaluating the early effects of temozolomide (TMZ) chemotherapy in low-grade glioma patients have been explored in [3].

Reference[4] presents a novel whole-brain diffusion tensor imaging (DTI) segmentation to delineate tumor volumes of interest (VOIs) for subsequent classification of tumor type. It uses isotropic and anisotropic components of the diffusion tensor to segment regions with similar diffusion characteristics.

In the findings of the paper[5], DTI is the only approach available to track brain white matter fibers noninvasively .

III. RATIONALE OF THE STUDY

For effective brain tumor treatment, an accurate identification of boundaries between tumor, edema and healthy tissue is critical in brain tumor patients. This is very challenging mainly owing to the fact that high-grade tumors are inherently diffuse and infiltrative and invade the surrounding healthy tissue.

Tumors are heterogeneous, comprising enhancing and non-enhancing tumor tissue types and edema, rendering the transition from tumor to healthy tissue gradual. It is therefore challenging, if possible at all, to identify a clear transition from healthy tissue to edema to tumor by an inspection of the conventional MR images.

Differentiation of tumor recurrence from treatment-related changes may be difficult with conventional MR imaging when newly enhancing lesions appear. This leaves recurrent brain tumor tissues untreated, likely leading to faster tumor spread and lowers the chance of survival. The study aims to differentiate tumor from the healthy brain tissues and the edema using Diffusion Tensor brain Images in brain tumor patients.

IV. PROPOSED WORK

The sequence of operations for identification of boundaries between tumor, edema and healthy tissue in diffusion tensor MRI images of brain contain various steps like image preprocessing and enhancement, image segmentation,feature extraction and classification.

a. Image preprocessing and enhancement

Before detecting the tumor in the image, preprocessing is done for increasing the reliability of optical inspection. Initially the DTI images of brain are acquired and they are pre-processed in order to extract the necessary information. During the process of image formation, the quality of images may degrade due to variety of causes such as out of focus, presence of noise, distortion of optical systems, the relative motion between the camera and the scene etc. Image is converted from RGB to grayscale mode. The grayscale is then enhanced to increase the quality of the image by applying median filter. Plot the histogram to study the strength of the pixels.

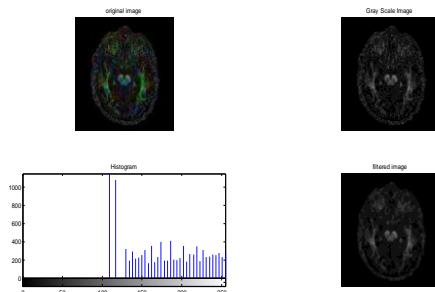


Fig 1: Image after preprocessing

b. Image Segmentation

Segmentation is the process of dividing an image into regions with similar properties such as gray level, color, texture, brightness, and contrast. The main goal in brain diffusion tensor MRI segmentation is to segment gray matter, white matter and cerebrospinal fluid. Segmentation is also used to find out the regions corresponding to tumors, edema, and other pathologies. The aim of medical image segmentation is to study anatomical structure, identify Region of Interest(ROI) i.e. locate tumor, measure tissue volume to measure growth of tumor and helps in treatment planning prior to radiation therapy.

Automatic segmentation of medical images is a difficult task as medical images are complex in nature and rarely have any simple linear feature. Although a number of algorithms have been proposed in the field of medical image segmentation, medical image segmentation continues to be a complex and challenging problem.

1. Artificial Intelligence Tools for Segmentation and Classification

Automatic segmentation methods have been based on artificial intelligence (AI) based techniques. AI techniques can be classified as supervised and unsupervised. Supervised segmentation requires operator interaction throughout the segmentation process whereas unsupervised methods generally require operator involvement only after segmentation is complete. Unsupervised methods are preferred to ensure a reproducible result; however, operator interaction is still

required for error correction in the event of an inadequate result.

1.1 Supervised method

In the supervised category, Artificial Neural Network (ANN) based algorithms are mostly used. ANN is composed of large number of interconnected processing elements (artificial neurons) working in unison to solve specific problems.

The main advantages of ANN are:

- ability to learn adaptively, using training data to solve complex problems.
- it can create its own organization depending upon the information it receives during learning time
- capability of performance in real time because of parallel configuration

1.2 Unsupervised Method

Most of the unsupervised algorithms are cluster based and not dependent on training and training data. The two commonly used algorithms for clustering are K-mean or Hard C-mean and Fuzzy C-means. K-means algorithm produces results that correspond to hard segmentation while fuzzy C-mean produces soft segmentation which can be converted into hard segmentation by allowing the pixels to have membership of cluster in which they have maximum value of membership coefficients.

1.2.1 Fuzzy c-Means algorithm

Clustering is the process of finding natural grouping clusters in multidimensional feature space. It is difficult because clusters of different shapes and sizes can occur in multidimensional feature space. A number of functional definitions of clusters have been proposed. Patterns within a cluster are more similar to each other than patterns belonging to different clusters. Image segmentation may be considered a clustering process in which the pixels are classified into the attribute regions based on the texture feature vector calculated around the pixel local neighbourhood. The Fuzzy c-Means algorithm is a clustering algorithm where each item may belong to more than one group (hence the word *fuzzy*), where the degree of membership for each item is given by a probability distribution over the clusters.

Since the absolute membership is not calculated, FCM can be extremely fast because the number of iterations required to achieve a specific clustering exercise

corresponds to the required accuracy. In each iteration of the FCM algorithm, the following objective function JJ is minimised:

$$J = \sum_{i=1}^N \sum_{j=1}^C \delta_{ij} \|x_i - c_j\|^2$$

Here, N is the number of data points, C is the number of clusters required, c_j is the centre vector for cluster j, and δ_{ij} is the degree of membership for the ith data point x_i in cluster j. The norm $\|x_i - c_j\|$ measures the similarity (or closeness) of the data point x_i to the centre vector c_j of cluster j. In each iteration, the algorithm maintains a centre vector for each of the clusters. These data-points are calculated as the weighted average of the data-points, where the weights are given by the degrees of membership.

For a given data point x_i , the degree of its membership to cluster j is calculated as follows:

$$\delta_{ij} = \frac{1}{\sum_{k=1}^C (\|x_i - c_k\|)^{2/m-1}}$$

where, m is the fuzziness coefficient and the centre vector c_j is calculated as follows:

$$c_j = \frac{\sum_{i=1}^N \delta_{ij}^m \cdot x_i}{\sum_{i=1}^N \delta_{ij}^m}$$

where δ_{ij} is the value of the degree of membership calculated in the previous iteration. Note that at the start of the algorithm, the degree of membership for data point i to cluster j is initialised with a random value θ_{ij} , $0 \leq \theta_{ij} \leq 1$.

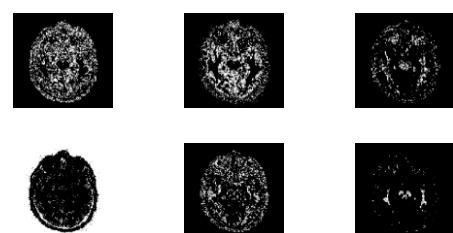


Fig 2: Fuzzy C-Means clustered image(6 clusters)

V. CONCLUSION

In this paper an algorithm using Matlab has been devised for the segmentation of brain tumor from DTI brain scanned images based on a range of operations like pre-

processing, Fuzzy C-means. The effectiveness of FCM is comparatively better than K means algorithm for overlapped datasets. In future, this system can be implemented with some other algorithm which will give more accuracy and save more time.

REFERENCES

- [1] Sternberg EL, et al., “ Utility of Diffusion Tensor Imaging in Evaluation of the Peritumoral Region in Patients with Primary and Metastatic Brain Tumors”, Americal Journal for Neuroradiology, 2014, Volume 35, Issue 3, pp:439 -444
- [2] Hongmin Cai, et al.,“Probabilistic Segmentation Of Brain Tumors Based On Multi-Modality Magnetic Resonance Images” in Proceedings of the IEEE International Symposium on Biomedical Imaging: from Nano to Macro, 2007.
- [3] Antonella Castellano, et al.,“Evaluation of low-grade glioma structural changes after chemotherapy using DTI-based histogram analysis and functional diffusion maps” European Radiology, 2016, Volume 26, Issue 5, pp: 1263-1273.
- [4] Timothy L. Jones, Tieman J. Bymes, et al.,“Brain tumor classification using the diffusion tensor image segmentation (D-SEG) technique” Neuro-oncology, 2015, Volume 17, Issue 3, pp: 466-476.
- [5] Denis Le Bihan, Jean-Francois Mangin, “Diffusion Tensor Imaging : Concepts and Applications” Journal of Magnetic Resonance Imaging, 2001, Volume 13, pp:534-546
- [6] Mori Susumu and J Donald Tournier, Introduction to diffusion tensor imaging and higher order models (Second edition) Academic Press, Oxford, UK, 2014.
- [7] B. H. Menze, M. Reyes, K. Van Leemput, et al., “ The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)” *IEEE Transactional Medical Imaging*, 2015, Volume 34, Issue 10, pp: 1993-2024.
- [8] Rafael C. Gonzalez and Richard E.Woods. Digital Image Processing (Third Edition) Dorling Kindersley Pvt. Ltd., India, 2014.
- [9] Cha, et al.,“Update on Brain Tumor Imaging: From Anatomy to Physiology” Americal Journal for Neuroradiology, 2006, Volume 27, pp: 475-487.
- [10] Marc C. Mabray, Ramon F. Bajaras and Soonmee Cha, “Modern Brain Tumor Imaging” Brain Research and Treatment Journal, 2015, Volume 3, Issue 1, pp:8- 23.
- [11] Timothy L. Jones, Tiernan J. Bymes, Guang Yang, Franklyn A. Howe, B. Anthony Bell Thomas R. Barrick “Brain Tumor Classification using the Diffusion Tensor Image Segmentation (D-Seg) Technique” Neuro-Oncology, Vol 17, Issue 3, March 2015, pp. 466-476.

Partial Discharges using Variable Frequency PRPDA Technique

M. Zubair Bhayo₁, M.Ali₂, Kalsoom Bhagat₃, Abdul Hameed₄

^{a,b,c} Department of Electrical Engineering, Mehran UET SZAB Campus Khairpur Mir's,

^d School of Automation, Northwestern Polytechnical University Xian China.

Abstract—At the applied voltage a disc-shaped cavity with partial discharges are measured at variable frequency (0.01-50 Hz). By varying the frequency it was observed that measured PD phase, magnitude of distributions and number of PDs per voltage cycles are varied. In the cavity, sequence of Partial discharge is simulated dynamically. For that purpose a model is presented with charge consistent. Simulated results shows that cavity surface and emission properties are effected by varying the magnitude of applied frequency, mainly conductivity of surface. This paper is illustrating the frequency dependence of PD in a cavity. The paper illustrates how the applied voltage amplitude and the cavity size can influence the frequency dependence PD activity.

Keywords: Partial discharges, simulation, modeling, variable frequency, cavities, disc-shaped and epoxy resin.

I. INTRODUCTION

Insulation is a major portion of the high voltage equipment. Numerous forms of insulating materials are used in high voltage electrical power system. The property of insulating material deteriorates enormously with the effect of partial discharge. Therefore, to keep the high voltage power equipment in healthy condition and to ensure the reliability of the power system the detection of partial discharge measurement is very important. Partial discharge occurs due to the defects in insulation system. Partial discharge results only if the electric field in the defects exceeds the thresh-hold field. The partial discharges are of different types e.g surface, internal and corona discharge. Partial discharge can occur in all medium of insulations due to presence of voids, cavities, gas bubbles in insulation. Sharp edges in insulation are one of the major sources of partial discharge (PD). For the detection and measurement of partial discharge in electrical insulation, variable frequency phase resolved partial discharge analysis is used [1, 2].

The main objective of measurement and detection partial discharge using different frequency is to decrease the size and power of the supply equipment. Though the power frequency greater than normal frequency is too important [3]. It has been observed that with increasing applied frequency the apparent charge per cycle of the applied voltage is decreased. Nowadays partial discharge activity in high voltage equipment is measured and detected using different voltage frequencies to ensure the reliability of power system. Recent studies [1, 5] have been used for comparing partial discharge measurement results obtained with different methods to analyze partial discharge behavior. Different diagnostic techniques have been adopted for understanding physical condition of the power cables. To detect partial discharges in the insulation of the service aged equipment, various voltage sources and oscillating waves in the range of 50-1000 Hz were used. In past literature experiments on power cables [1,4] and of generators stator [6] were performed and it was found that the partial discharge quantities such as partial discharge extinction voltage, PD level have no fundamental differences obtained due to oscillating voltage waves and 50(60)Hz ac energizing methods .

The field of diagnostics started 1992 at KTH with a project to diagnose Cross-linked Polyethylene cables that suffered by water-trees [7, 8]. The method applied was High Voltage Dielectric Spectroscopy [12] performed in the low frequency range, i.e. from 1 mHz to 1000 Hz. The method of dielectric spectroscopy has also been applied on oil-paper insulated high voltage equipment such as power transformers [10] and oil impregnated paper cables [11]. Partial discharges are measured and analyzed at different frequencies of the applied voltage through the recent developed technique of (VF_PRPDA) [13,14]. At different applied frequencies, the local conditions at defects have been changed due to the frequency dependence

of the partial discharges. From local conditions at defects which is due to the frequency dependence can be utilized for insulation diagnostic purpose.

II. LITERATURE REVIEW

Gafvertl et al have used phase resolved partial discharge variable frequency analysis technique for partial discharge measurements. The variable frequencies were used in place of normal power frequency. It is observed how the frequency dependency is inclined because of the conductivity through insulation and cavity walls. The results obtained with simulation have been compared with measurement results on mica insulated stator bar. We can suggest an analysis on mica insulated bar stator bar through comparing with simulation results. The authors have concluded that with type of modeling they used in this paper are useful to understand variable frequency partial discharge forms [1]. G. Chen et al have also worked on the partial discharge (PD) under variable frequency. However they have focused either on spherical or ellipsoidal as the best common forms of cavities establish in insulation material are of this type. The model presented in this paper can be utilized in a homogenous dielectric to describe the spherical cavity. The developed model is useful to determine the effect of applied frequency on partial discharge action. The authors have concluded that statical time lag and dielectric material time constant play an important role on partial discharge in spherical cavity [2].

Bodega et al have described a method to simulate partial discharge measurements in bounded spherical cavities under the range of 0.1Hz-1000Hz. From the simulation results the information about the impact of frequency on PD process have been derived. They have investigated from both the experimental results and numerical analysis, the voltage frequency have large impact on the Partial discharge behavior. The authors have successfully simulated the partial discharge process with use of numerical analysis based on mathematical model [3]. Morshuis et al have investigated the insulation of the service aged components with the help of very low frequency (VLF).Based on this study the authors have concluded that PD level obtained with above than 200Hz is slightly small than that observed with the 50Hz ac energizing method. However the PD process at lower frequencies can be either very close to that at 50Hz or quite different [4].

III. RESEARCH METHODOLOGY

A. MODEL

Simple test objects are used for measurement of PD under variable frequencies by applying voltage as in figure 1.1. One cylindrical cavity is present in test object in a exfoliate insulation. Insulating plates are passed together in between two electrodes of disc shaped to create the test objects in this way. Brass and cast in epoxy is used to make electrodes in order to prevent discharges at the edges of electrodes. To drill a hole in one or more insulating plate for creating cylindrical cavity. Advantage of using this type of cavity shape is to make it more accurate and easy. And no PD seen in between insulating plates up to 12 kV applied voltage.

Drilled hole diameter changed for varying cavity diameter. By changing insulating plates thickness height of cavity is to be changed.

Rearranging the insulating plates between electrodes, cavity location is to be changed. To create cavities height about 1 mm thickness of insulating plates are taken. In practice cavities are less thick i-e machine insulation. As 1 mm plates are easy in use.

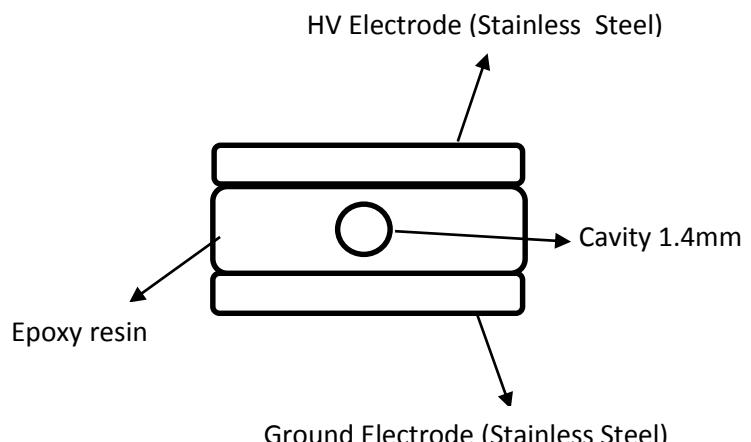
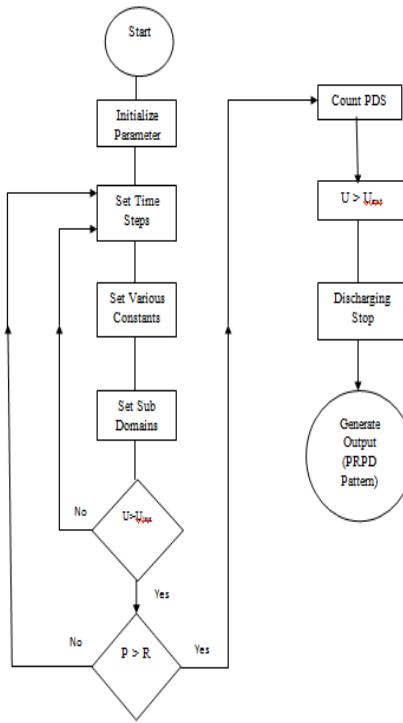


Figure 1: Test object with cylindrical cavity. Cross-section along symmetry axis.

B. FLOW CHART

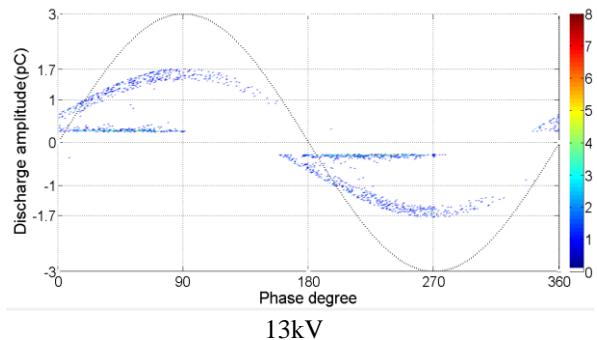


First of all model has been designed in the matlab software. We initialize all the parameters in the designed model. We set the time steps to which model will run. After selecting the time steps, the model geometry parameters have been set. In the next step of model will check the condition that void voltage should be greater than the inception voltage. If the condition is true the program will move to the next condition of probability of electrons should be greater than random number. If both conditions are true then the program will move forward on the other hand if the condition is false the program will back to the time steps to check the condition again. Then the model will count number of partial discharges occurred. After counting number of partial discharges, if void voltage is greater than extinction voltage the discharge will stop. Finally the matlab model will generate the output in the form of phase resolved partial discharge pattern.

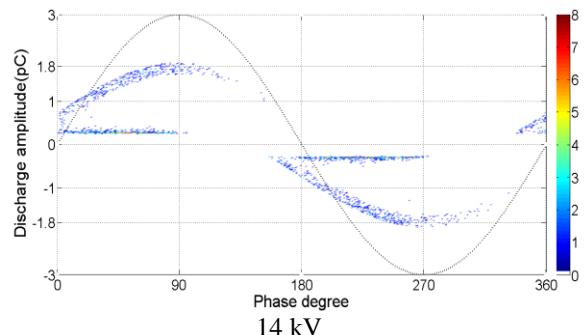
IV SIMULATION RESULTS

a) PRPD PATTERN UNDER VARIOUS APPLIED VOLTAGE AMPLITUDE:

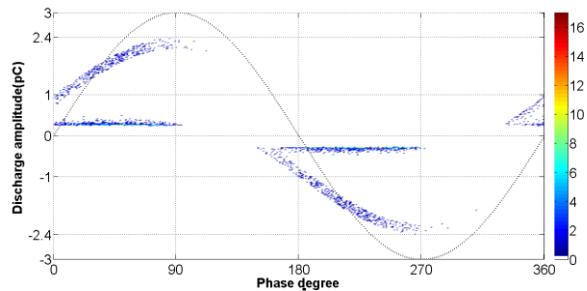
This section illustrates how the PD frequency dependence is influenced by amplitude of the applied voltage. With a diameter of 1.5 mm in an insulated cavity the PD activity is simulated at applied voltage amplitude of 13, 14, 15, 16, 17 and 18 kV. Below the voltage amplitude of 8kV, there were no PDs observed.



13kV



14 kV



15 kV

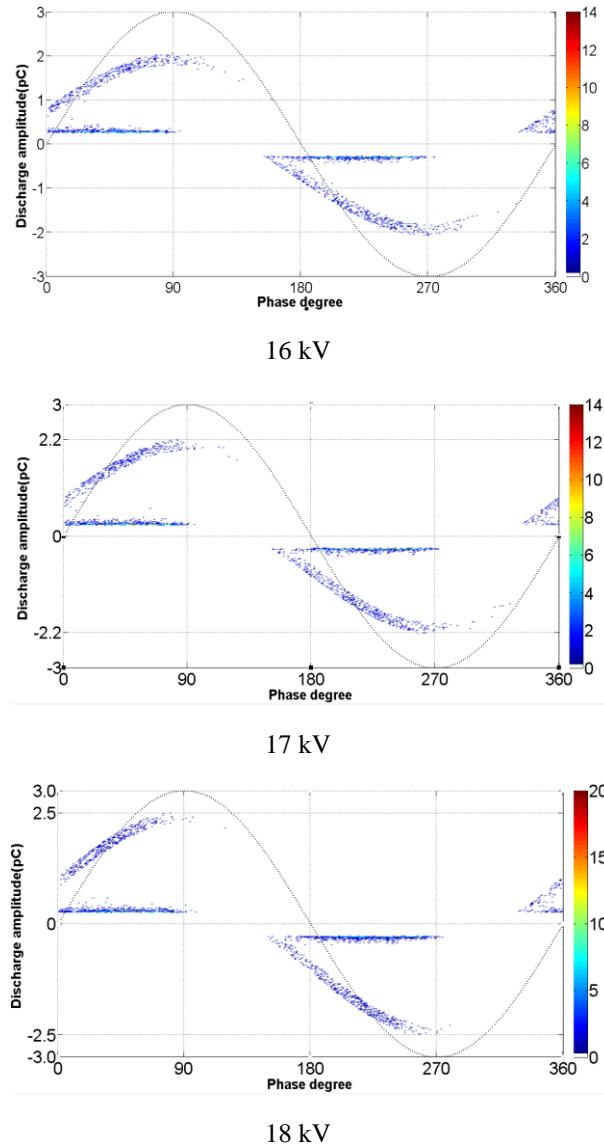


Figure 2: PRPD pattern under various voltages amplitude.

Figure 2 shows the simulated PD patterns at amplitude 13, 14, 15, 16, 17 and 18kV at power frequency with diameter of 1.5mm. The PD activity in the cavity clearly changes with the voltage Amplitude. As the voltage is increased from 13kv-18kv, the maximum PD magnitude increase from about 1700 pC at 13kv to about 2500 pC at 18kv. In contrast the minimum PD magnitude is approximately constant. Hence, there is a wider spread in PD magnitude at higher voltage than at lower voltage.

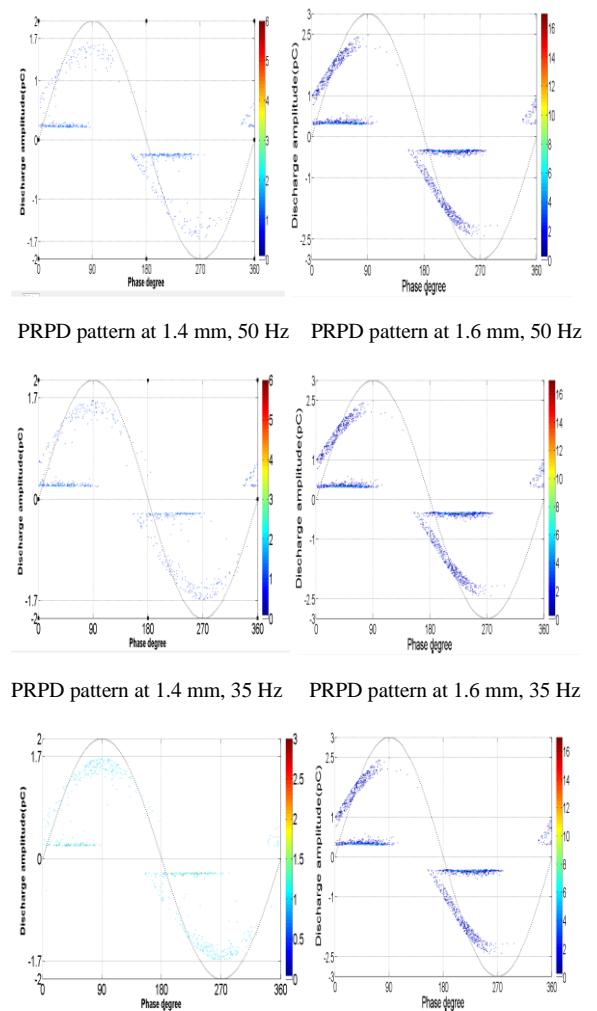
The PD activity is similar with respect to the polarity of applied voltage for an insulated cavity. With the increasing frequency the statical effect was observed with the change in maximum PD magnitude. There was a wide spread in PD magnitude at 50 Hz

frequency. The average phase positions of positive PDs at 0.01 Hz and 50 Hz are 105 deg and 140 deg, respectively.

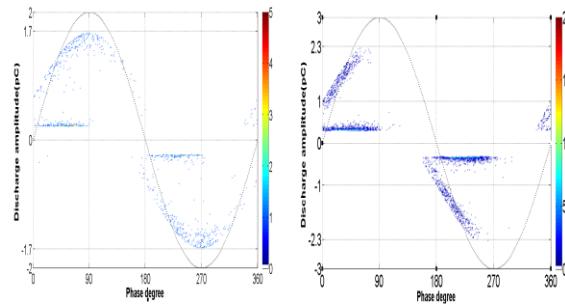
b) PRPD PATTERN UNDER DIFFERENT CAVITY SIZE:

In this section it is described that the cavity diameter can also influence the partial discharge frequency dependence of the test object. With a 15 kV applied voltage amplitude and different frequency, the partial discharge activity for insulated cavities of 1.4 mm and 1.6 mm were simulated respectively.

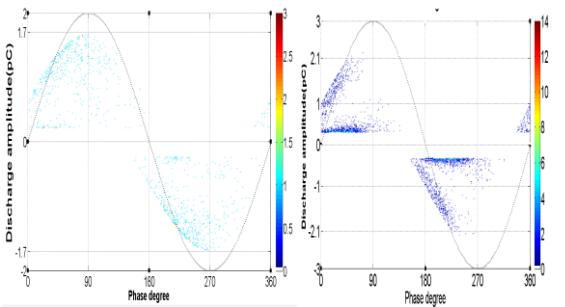
10 kV and 7 kV were the lowest voltage amplitude where partial discharges were observed with a cavity diameter of 1.4 mm and 1.6 mm respectively. It is observed that with a cavity diameter of 1.4 mm, the field enhancement is lower than the cavity diameter of 1.6 mm. Because of that, the partial discharge inception voltage was large with 1.4 mm diameter than 1.6 mm.



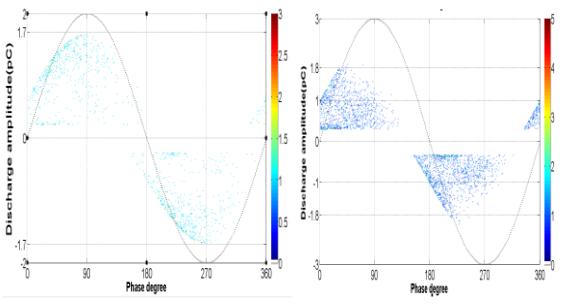
PRPD pattern at 1.4 mm, 25 Hz PRPD pattern at 1.6 mm, 25 Hz



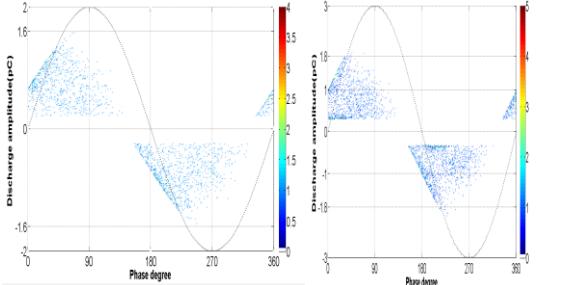
PRPD pattern at 1.4 mm, 15 Hz PRPD pattern at 1.6 mm, 15 Hz



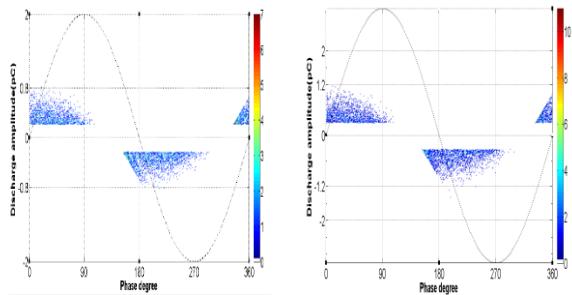
PRPD pattern at 1.4 mm, 10 Hz PRPD pattern at 1.6 mm, 10 Hz



PRPD pattern at 1.4 mm, 5 Hz PRPD pattern at 1.6 mm, 5 Hz



PRPD pattern at 1.4 mm, 1 Hz PRPD pattern at 1.4 mm, 1 Hz



PRPD pattern at 1.4 mm, 0.1 Hz RPD pattern at 1.6 mm, 0.1 Hz

Figure 3: Comparison of PRPD pattern at 1.4 mm and 1.6 mm void diameter 50-0.1 Hz frequencies.

For test objects with varying cavity diameter a comparison of the PD magnitude at a frequency 50-0.1 Hz have been illustrated in Figure. 3. It was observed, there were different partial discharge inception voltages at different cavity diameters of the test objects. The PDs were concentrated to one or more distinct magnitudes for all cavity diameters. This illustrates that the PD activity does not influenced by the statical time lag significantly.

It was also observed that there were concentration of PDs at 0.01 Hz to a magnitude of about 800 pC for all diameters. There was a smaller concentration of PDs above 1200 pC for a diameter 1.6 mm. there was one preferred magnitude for a diameter of 1.4 mm at 50 Hz. It was too noticed that with increasing frequency the actual number of PDs per cycle are higher and increased, since the no. of similar detected partial discharges increased. The test objects with different cavity diameter have difference in PD frequency dependence was due to that the no. of PDs per cycle, whereas magnitude of PD were same.

It was also found that the no. of PDs per cycle was about the same for the smallest cavity diameter at all frequencies. In addition to that the no. of PDs per cycle and the maximum no. of similar detected PDs decreased with decreasing frequency for the largest cavities. Finally it was discovered that test objects with largest cavity diameter have higher no. of PDs because of PD inception voltages decreased with increasing cavity diameter. Also with increasing cavity diameter the no. of similar PD detected increased.

V CONCLUSION

More information about the condition of an insulation system is achieved from simulation of partial discharges at variable frequencies of the applied voltage than from simulation at the single frequency. It becomes possible to differentiate between insulated

cylindrical cavities of different heights and cavities bounded by an electrode through simulation of partial discharge at variable applied frequency.

Using two dimensional field model, the sequence of partial discharges in insulated cylindrical cavity at different frequencies were simulated. To obtain consistent charge densities and currents in the model the discharges were simulated dynamically. From the two dielectric time constants and statical time lag, the model can be effectively utilized to predict influence on the partial discharge frequency dependence. In this type of modeling, the simulation time is a critical parameter used.

ACKNOWLEDGEMENT

The authors are thankful to Mehran University of Engineering & Technology SZAB Campus Khairpur Mir's, Sindh, Pakistan, for providing all necessary laboratory facilities for the completion of this research work. The authors also would like to thank, Prof. Agha Zafarullah Pathan, Faculty member, Mehran University of Engineering & Technology SZAB Campus Khairpur Mir's, Sindh, Pakistan, and Professor Dr. Mukhtiar Ahmed Unar, Faculty Member, MUET Jamshoro, Sindh, Pakistan for their valuable input to this paper.

REFERENCES

- [1] Mirza Batalovic, Kemo Sokolija, Mesud Hadzialic and Nejra Batalovic, "Partial Discharges and IEC Standards 60840 and 62067: Simulation Support To Encourage Changes, Tehnicki vjesnik, Vol. 23, Issue 2, pp: 589-598, 2016.
- [2] Pragati Sharma and Arti Bhandakkar, "Simulation Model of Partial Discharge in Power Equipment", International Journal of Electrical and Electronics Research, ISSN 2348-6988 (online), Vol. 3, Issue 1, pp: 149-155, Month : January – March-2015.
- [3] S.D.M.S Gunawardana, A.A.T. Kanchana, P.M. Wijesingha, H.A.P.B. Perera, R. Samarasinghe and J.R. Lucas, "A Matlab Simulink Model for a Partial Discharge Measuring System", Electrical Engineering Conference [EECon], 2015.
- [4] Engr. Manzoor Ahmad Khan and Dr. Amjadullah, "Measurement Of Partial Discharge (PD) In High Voltage Power Equipment", First International Conference on Emerging Trends in Engineering, Management and Sciences", pp: 28-30, December-2014 (ICETEMS 2014) Peshawar, Pakistan.
- [5] M. S. Hapeez, A. F. Abidin, H. Hashim, M. K. Hamzah and N. R. Hamzah, vol. 24, no. 1, April 2011, 41-55,"Analysis and classification of different types of Partial Discharges by Harmonic Orders", Elektronika Ir Elektrotehnika, ISSN 1392-1215, Vol. 19, Issue 9, 2013.
- [6] Nidhi SinghSubhra Debdas and Rajeev Chauhan, "Simulation & Experimental Study Of Partial Discharge In Insulating Materials For High Voltage Power Equipments", International Journal of Scientific & Engineering Research, ISSN 2229-5518, Vol. 4, Issue 12, December-2013.
- [7] Y. Z. Arief, W. A. Izzati and Z. Adzis, "Modeling of Partial Discharge Mechanisms in Solid Dielectric Material", International Journal of Engineering and Innovative Technology (IJEIT) Vol. 1, Issue 4, April-2012.
- [8] N.O. Ogbogu and A. Akhikpemelo, "Partial Discharge Mechanism Modeling Of Solid Dielectrics", Continental J. Engineering Sciences, pp: 14 – 21, Vol. 7, Issue 3, 2012.
- [9] Asima Sabat and S. Karmakar, "Simulation of Partial Discharge in High Voltage Power Equipment", International Journal on Electrical Engineering and Informatics, Vol. 3, Issue 2, 2011.
- [10] Kheng Jern Khor, "Partial Discharge Propagation and Sensing in Overhead Power Distribution Lines", School of Electrical and Computer Engineering College of Science, Engineering & Health RMIT University, March-2010.
- [11] R. Neimanis, On Estimation of Moisture Content in Mass Impregnated Distribution Cables, PhD thesis, KTH 2001, TRITA-EEK-0101, ISSN 1100-1593.
- [12] Nenad Kartalovic, Dragan Kovacevic, and Srdjan Milosavljevic, "An Advanced Model of Partial Discharge in Electrical Insulation", FACTA UNIVERSITATIS (NI'S) SER.: ELEC. ENERG.
- [13] B. Holmgren, Dielectric response breakdown strength and water-tree content of medium voltage XLPE cables, Licentiate thesis, KTH 1997, TRITA-EEA-9705, ISSN 1100-1593.
- [14] P. Tharning, Water tree Dielectric spectroscopy, Licentiate thesis, KTH 1997, TRITA-EEA-9703, ISSN 1100-1593.
- [15] H.Edin, Partial Discharges studied with variable frequency of the applied voltage. PhD Thesis, KTH, Stockholm 2001.
- [16] j.Giddens, H.Edin and U.Gafvert, "Measuring system for Phase Resolved Partial Discharge Detection at low frequency", 11thInt.symp. on High Voltage Engineering, London, UK, 1999, pp 5.288-5.231.
- [17] George Chen and Fauzan Baharudin, "Partial discharge modelling based on a cylindrical model in solid dielectrics", 2008 International Conference on Condition Monitoring and Diagnosis, Beijing, China, pp: 21-24, April-2008.

Vehicle Power Line Channel Modelling under CST Microwave Studio

Mohammed Fattah

Transmission and Information Processing Team
Moulay Ismail University
Meknes, Morocco
fattahm@gmail.com

Abstract—For every car manufacturer, it is essential to reduce the weight of vehicles. To help manufacturers reduce weight factors, cable industries propose several solutions to reduce the weight of harnesses such as creating new wires with reduced conducting sections. However, we offer another complementary solution, based on using the PLC (Power Line Communication) technology inside a vehicle. In this article, we address the problem of vehicle PLC, focusing on the channel modelling. The aim of this paper is to design a novel PLC channel model under CST MWS (CST Microwave Studio) software, able to emulate a real vehicle PLC environment. Simulation results are compared to experimental measurements.

Keywords- *Power Line Communication; Vehicle Power Line Channel; Channel Modelling; 3D Model*

I. INTRODUCTION

Vehicle manufactures continue to increase the electrical and electronic devices in vehicles. This growing trend comes from the demand for electronic control and command equipment. All those equipments offer to the driver comfort and safety. Many techniques are being developed: anti-collision radar, reversing radar, white line crossing detection, space measuring system, indication of a vehicle at a blind angle, adjusted the speed according to the space behind the vehicle in front, and triggers the brakes in an emergency etc. With increased numbers of equipment, the weight of vehicles has been increasing (20 kg per year in Europe). But in the same time, Euro 6 standards [1] require a reduction pollutant emission (80 mg/km). The weight of the car is related directly to 75% of fuel consumption. This means that reducing the weight of a vehicle by 100 kg is equivalent to reducing the fuel economy by 11 g of CO₂ per kilometre. That is why manufacturers have set themselves the goal of quickly reducing the weight of the vehicle by 200 kg by the year 2020.

All these onboard systems mean an increase in the number of wires and cables in the vehicle. In this context, our solution is based by using the PLC technology. The PLC technology is a communication method that uses electrical wiring to simultaneously carry both data and electric power.

The aim of this paper is to design a vehicle PLC channel model under CST MWS software [2], which could be reliable in analyzing broadband vehicle PLC systems. This modelling study will be validated by experimental measurements.

S. Mazer, M. El Bekkali, R. Ouremchi, M. El Ghazi

Transmission and Information Processing Laboratory
Sidi Mohamed Ben Abdellah University
Fez, Morocco
fattahm@gmail.com

The paper is organized as follows. In Sections II, we describe the vehicle PLC channel model under CST MW Studio. In Section III, we provide the simulation results and we compare it with experimental results. Finally, the conclusions follow in Section IV.

II. VEHICLE CHANNEL MODELLING

The frequency response of the current vehicle network electrical is not flat but has resonances and fading, due to echoes and reflections between the transmitter and the receiver [3].

There are many reasons behind these changes for example, the coupling between the different wires of the strand, the wiring ohmic losses and the multipath characteristics of the channel.

Also, the increasing number of interconnections in the car is inevitable despite the use of multiplexing. It causes disadvantages in conception and fabrication that makes the fault diagnosis and detection very difficult. This implies that the classic electric network of the vehicle is currently reaching their technological limits. It is therefore necessary to design a new on-board electrical network of a vehicle by designing a new cabling architecture.

Our idea is to reduce the number of wires to two conductors plus the ground. That is means; the channel transmission is composed of two cables, the first cable carrying the power to supply the electronic and electrical components, and the second cable for data transmission as show in Fig. 1.

In Fig. 1 the battery is replaced by the Randles battery model [4] as show in Fig. 2.

Fig. 3 shows the cable used in the experimental measurements. The cable is a multi-core, round twin cable constructed from two cores of thin wall cable with a PVC (Poly Vinyl Chloride) outer sheath. It is suitable for low voltage car. Thin wall cable is the most used as standard by car manufacturers due to its high-performance characteristics. The cable specifications are listed in Table I.

The Fig. 4 shows a section of a vehicle electrical network with four outlets, E, S, P1 and P2. This section is used in measurements to determine the S-parameters between E

(input) and S (output) in the band [500 KHz – 70 MHz]. The modelling of the vehicle PLC channel (Fig. 4) by using the MLT approach (Multiconductor Transmission Line) under Matlab software and by using scattering matrix [S] under ADS software (Advanced Design System) was validated and published in [5][6].

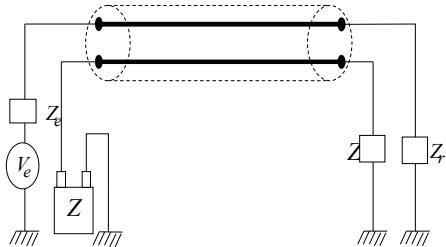


Fig. 1. Transmission channel with two conductors

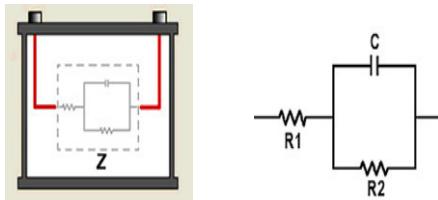


Fig. 2. Randles battery model

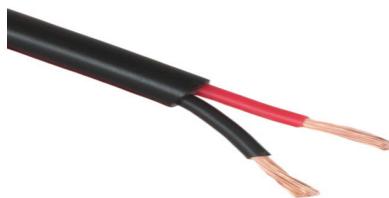


Fig. 3. Cable used in measurements

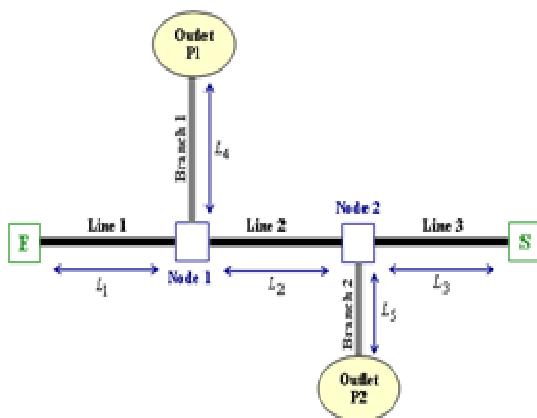


Fig. 4. Section of a vehicle electrical network

Parameter	Specification
Cores manufactured	ISO 6722-1:2011(Class B)
Voltage rating	12 V & 24 V
Nominal current rating	2x25 A
No./size of conductors	2x28 /0.30 mm
Conductor cross section	2x2.0 mm ²
Overall cable diameter	7.8 mm
Conductor material	Plain copper
Core insulation material	PVC (hard grade)
Sheath insulation material	PVC
Working temperature	-15 to +70°C

In
this

paper, we are interested in modelling the vehicle PLC channel in 3D simulation under CST MWS by using FIT method (Finite Integration Technique).

The FIT was first proposed in 1977 by Thomas Weiland [7]. The FIT is a discretization scheme for Maxwell's equations in their integral form suitable for computers and it allows to simulate real-world electromagnetic field problems with complex geometries [8]. In [8], Algebraic properties of the discrete formulation make it possible to develop long-term stable numerical time integration. And, the FIT is a generalization of the FDTD method (Finite Difference Time Domain) [9].

Also, CST MWS software is based on FIT method. The power cable structure is built in 3D after the definition of calculation volume. 3D simulation can take into consideration all the electromagnetic phenomena resulting of the propagation of broadband signals in power cable. The structure is excited by a waveguide port. Fig. 5 shows the 3D structure of the Cable by using CST MWS, respecting the same specifications of the cable used in the measurements (Table I).

The frequency response of an electric cable depends on the geometrical parameters: conductor cross section, overall cable diameter, distance between conductors, number of conductors, thickness of insulators, shape and length of the cable. It also depends on the technological parameters: relative permittivity of the insulators, loss angle of the insulators and electrical conductivity of the conductors. The parameters are presented in Table II. Also, it should be noted that the parameters (relative permittivity and loss angle of the insulators) are unknown when designing the cable. Their values were determined by using 3D simulation in order to get as close as possible to the measurement.

Fig. 6 shows the vehicle electrical network segment modelled under CST MWS. It corresponds to the vehicle electrical network segment used in the measurements (Fig. 4). The battery is represented by its equivalent impedance described by the Randles battery model.

TABLE I. SPECIFICATIONS CABLE USED IN THE MEASUREMENTS

TABLE II. GEOMETRIC AND TECHNOLOGICAL PARAMETERS OF THE CABLE

Dimension	Value
A	3.9 (mm)
B	1.5 (mm)
D	4 (mm)
R	0.79 (mm)
S	2 (mm^2)
ϵ_{int}	1 (mm)
σ	58.6 1E6 (S/M)
$\tan(\delta_{int})$	0.01
$\tan(\delta_{ext})$	0.03
ϵ_{int}	2.4
ϵ_{ext}	3.1

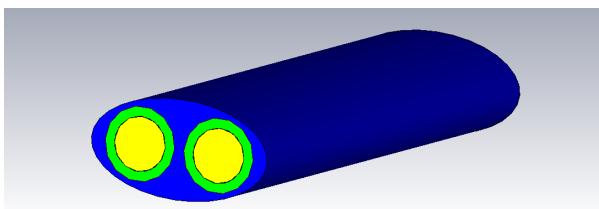


Fig. 5. 3D structure of Power Cable

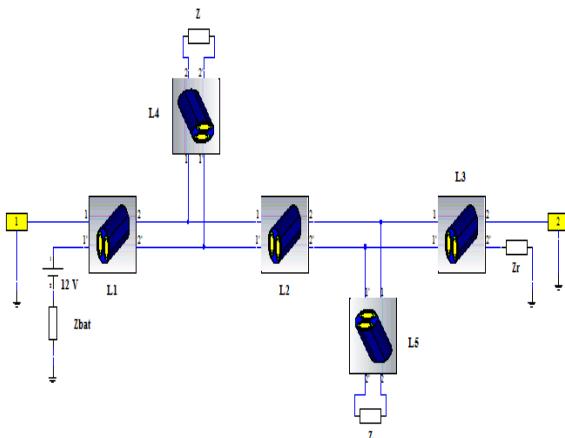


Fig. 6. Section of the vehicle electrical network modelled under CST MW

III. SIMULATION RESULTS AND DISCUSSION

In this section, we compare the simulation results with experimental measurements. The different simulations were obtained by the CST MWS software.

To calculate the S-parameters of the cable under CST MWS, we used the Time Domain solver. The mesh refinement of the structure is automatically done by the software, in order to achieve an accuracy of 1% on the obtained results. The simulated CST-model segment is shown in Fig. 6.

In 3D model, we consider the non-dispersive sheath material. Also, in this model, the skin effect is taken into account.

To validate this modelling method, we proceeded to a comparison between the simulation results and experimental measurements made on a model representing the automotive PLC used in the industry, for more information see [5]. We study the transfer function between E and S for two configurations P1 and P2 (Fig. 6):

- Configuration Open Circuit (OC): no devices are connected in the P1 and P2.
- Configuration Closed Circuit (CC): identical devices are connected to P1 and P2. To simulate this, the impedance matches the electrical and electronic equipment such as lamps, motors, sensors, etc... Measurements made by the manufacturers or suppliers show that the values of impedances can vary from 1Ω to $1\text{K}\Omega$ on a frequency band up to 70 MHz. Four impedance values were considered: $Z = 1 \Omega$, $Z = 50 \Omega$, $Z = 120 \Omega$ and $Z = 1\text{K}\Omega$.

The cable length used in this model are (Fig. 6):

$$L1 = L3 = 0.6 \text{ m}; L2 = 0.4 \text{ m}; L4 = L5 = 0.5 \text{ m}$$

The comparison between the simulation results and experimental measurements in the band [500 KHz – 70 MHz] is presented in Fig. 7 for OC configuration, Fig. 8 for CC configuration $Z = 1 \Omega$, Fig. 9 for CC configuration $Z = 50 \Omega$, Fig. 10 for CC configuration $Z = 120 \Omega$ and Fig. 11 for CC configuration $Z = 1\text{K}\Omega$.

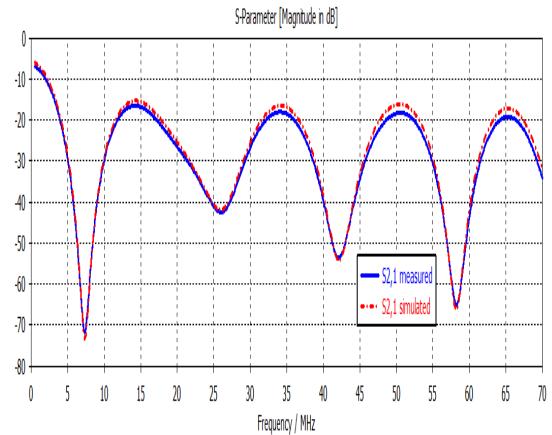


Fig. 7. S21 in OC configuration

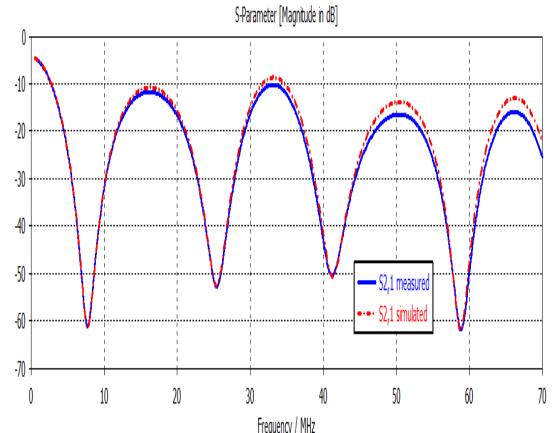


Fig. 8. S21 in CC configuration ($Z = 1 \Omega$)

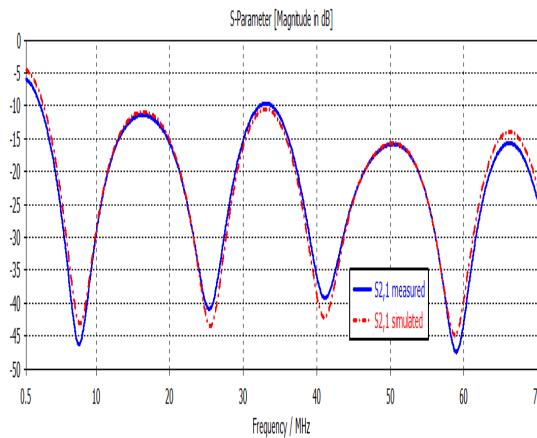


Fig. 9. S21 in CC configuration ($Z = 50 \Omega$)

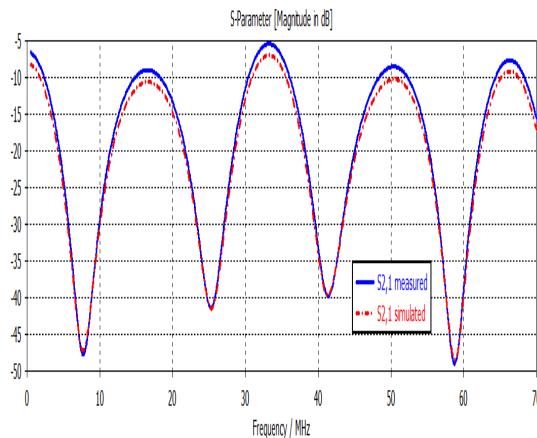


Fig. 10. S21 in CC configuration ($Z = 120 \Omega$)

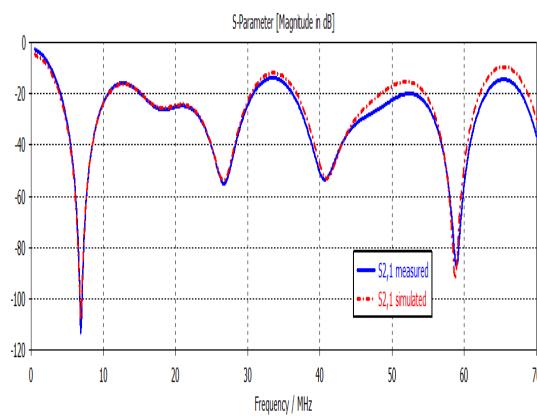


Fig. 11. S21 in CC configuration ($Z = 1 \text{ k}\Omega$)

For all configurations, we can observe four attenuation peaks: the first at 7.4 MHz, the second at 25.66 MHz, the third at 41.7 MHz and the fourth at 58.48 MHz. This phenomenon is mainly induced by a strong impedance mismatch at P1 and P2.

Also, we can see an average difference of 1.5 dB between measurements and simulations. This difference is due to the assumptions made in 3D modelling; assumes that the constant loss angle on the frequency band (non-dispersive sheath material).

In OC configuration, we find that attenuation is maximized (example: -74 dB for the 7.4 MHz frequency). This is caused by the connection of the derivation in open circuit. This derivation acts as a band-stop filter. It behaves like a short circuit at the connection points to the fading frequencies and reflects the incident wave towards the source. Moreover, the measurement confirms the simulation. In addition, it can be noted that the connection of an impedance (CC configuration) improves the frequency response (Fig. 8, Fig. 9 and Fig. 10).

finally, the small difference between measurements and simulations validates the CST-model in the band [500 kHz – 70 MHz].

In this paper, the PLC channel in the in-vehicle scenario has been discussed. A difference of $1 \text{ dB} \pm 0.5$ (97% confidence) is obtained between the measured and the simulated. Therefore, this model under CST MWS software reproduces approximately the same frequency behavior of the vehicle PLC channel.

In order to generalize the proposed model, this study needs to be extended to a more complex vehicle network.

Later, this model will be implemented in a software communication tool designed to optimize channel coding and modulation schemes.

REFERENCES

- [1] Euro 6 standards. http://europa.eu/rapid/press-release_IP-15-5945_en.htm
- [2] CST Microwave Studio software. <https://www.cst.com/products/cstmws>
- [3] M. O. Carrion, "Communications sur le réseau d'énergie électrique d'un véhicule : modélisation et analyse du canal de propagation," Ph.D. dissertation, Science and Technology Univ., Lille, French, July 2006.
- [4] B. Hariprakash, S. K. Martha, A. K. Shukla, Monitoring sealed automotive lead-acid batteries by sparse-impedance spectroscopy, J. Chem. Sci, Volume 115, (Issue 5 & 6), October–December 2003, Pages 465–472.
- [5] M. Fattah, R. Ouremchi, M. El Bekkali, and S. Mazer, Modelling the Channel Transfer Function of the Vehicle Power Line Channel, International Review on Modelling and Simulations (IREMOS), Volume 3, (Issue 6), December 2010, Pages 1273-1280.
- [6] S. Mazer, M. Fattah, M. El Bekkali and R. Ouremchi, Modeling the transfer function of the automotive PLC by using the S-Parameters, International Journal on Engineering Applications (IREA), Volume 1, (Issue 4), November 2013, Pages 259-262.
- [7] Thomas Weiland, A discretization method for the solution of Maxwell's equations for six-component fields, Electronics and Communications AEU, Volume 31, (Issue 3), Pages 116–120, February 1977.
- [8] M. Clemens and T. Weiland, Discrete Electromagnetism with the Finite Integration Technique, Computational Science and Engineering, Springer, Berlin, Heidelberg, volume 28, 2003, Pages 65–87.
- [9] Kane Ye, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, IEEE Transactions on Antennas and Propagation, Volume 14, (Issue 3), 14 (3), May 1966, Pages 302–307.

Big Data Clustering Model based on Fuzzy Gaussian

Amira M. El-Mandouh

Beni-Suef University

amiramohey@fcis.bsu.edu.eg

Laila A. Abd-Elmegid

Helwan University

drlaila_mohamed@yahoo.com

Hamdi A. Mahmoud

Beni-Suef University

dr_hamdimahmoud@yahoo.com

Mohamed H. Haggag

Helwan University

mohamed.haggag@fci.helwan.edu.eg

Abstract — Clustering is also known as data segmentation aims to partitions data set into groups, clusters, according to their similarity. Cluster analysis has been extensively studied in many researches. There are many algorithms for different types of clustering. These classical algorithms can't be applied on big data due to its distinct features. It is a challenge to apply the traditional techniques on large unstructured data. This study proposes a hybrid model to cluster big data using the famous traditional K-means clustering algorithm. The proposed model consists of three phases namely; Mapper phase, Clustering Phase and Reduce phase. The first phase uses map-reduce algorithm to split big data into small datasets. Whereas, the second phase implements the traditional clustering K-means algorithm on each of the spitted small data sets. The last phase is responsible of producing the general clusters output of the complete data set. Two functions, Mode and Fuzzy Gaussian, have been implemented and compared at the last phase to determine the most suitable one. The experimental study used four benchmark big data sets; Covtype, Covtype-2, Poker, and Poker-2. The results proved the efficiency of the proposed model in clustering big data using the traditional K-means algorithm. Also, the experiments show that the Fuzzy Gaussian function produces more accurate results than the traditional Mode function.

Keywords—Big Data; MapReduce; Fuzzy Gaussian; K-means.

1. INTRODUCTION

Data mining is a mechanism extracting the information from data. It is challenging to get relevant information and provide it within shortage time [4]. In data mining; supervised learning and unsupervised learning are the two learning approaches utilized to mine data [5]. In the Supervised learning; data includes both input and the desired outcome. The desired results are known and are given in inputs to the model during the learning procedure. The neural network, Multilayer perception, Decision tree are examples of supervised models. On the other hand in unsupervised learning. The desired outcome is not given to the model during the learning procedure. This method can be used to cluster the input data in classes by their statistical properties only. These models are for the various type of clustering, k-means, distances and normalization, self-organizing maps.

Data mining had some algorithms like classification, clustering, regression and association rule. Clustering is a task

to group data by their similarities and dissimilarities from data elements; mainly it is difficult at the time of big dataset. Clustering method converts that information into various clusters where the object in that group has similar properties as compared to other but not same to other clusters properties.

The rest of this article is planned as follows. Section II presents related works about the k-means algorithm. Part III contains brief discussion about algorithms which used a new model such as map-reduce, k-means, mod & Gaussian. The proposed algorithm is explained in Section IV. Section V introduces the experimental results by using four big data namely; Covtype, Covtype-2, Poker Hand and Poker Hand-2. In Section VI, we discuss the actual importance of the model in the conclusion section.

2. RELATED WORK

Clustering is a process for partitioning datasets. It is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities [2] [14]. This technique is helpful for an optimum solution. K-mean is the most famous clustering method. Mac Queen in 1967, firstly introduced this algorithm, though the idea went back to Hugo Steinhaus in 1957 [3].

Y. S. Thakare et al. [6] discussed the performance of k-means algorithm which is evaluated with various datasets such as "Iris," "Wine," "Vowel," "Ionosphere" and "Crude oil" dataSets and different distance metrics. It is assumed that performance of k-means clustering depends on the datasets has been used as distance metrics. The k means clustering algorithm is evaluated using recognition rate for a different number of the cluster. This work assisted in choosing suitable distance metric for an appropriate purpose.

SK Ahammad Fahad [7] proposed a method for making the algorithm which is consuming time effective and efficient for reduced complexity. The quality of their resulting clusters heavily depends on the selection of initial centroid and changes in data clusters in the subsequent iterations. After a definite number of iterations, a small part of the data points changes their clusters. Their approach; first gets the initial centroid and sets intervals between those data elements which will not exchange their cluster and those which may exchange their

cluster in the subsequence iterations. So, it will decrease significantly in case of large datasets.

Two methods for clustering the large datasets using MapReduce has presented in [8]. Firstly, "K-Means Hadoop MapReduce (KM-HMR)" which focused on the MapReduce implementation of regular K-means. The second one improves the clusters quality to create clusters with distances that are maximum in intra-cluster and minimum in inter-cluster for large datasets. The results of their introduced methodology present enhancement in the execution time efficiency of clustering. Experiments executed on original K-means, and proposed model shows that their approach is both powerful and efficient.

Mugdha et al. [20] introduce an approximate algorithm based on k-means. The algorithm minimizes the complexity measure of k-means by calculating over only those attributes which are of interest is proposed here. Their algorithm cannot manipulate categorical data completely until it is transformed it into equivalent numerical data. Manhattan distance concept has been practiced, which in turn decreases the runtime. It is a new method for big data analysis. Their algorithm is scalable, very fast, and have great accuracy. It succeeded to overcome the disadvantage of k-means of an uncertain number of full iterations. They set a fixed number of iterations, without losing the precision.

G. Venkatesh1 et al.[21] Present a method is called layers three aware traffic clustering based on parallel K-means and the distance metric for minimizing the network traffic cost. Their method applied map-reduce model in three layers. Various algorithms are discussed in their paper; they compared between it, e.g., "Bisecting K-Means", "K-Means Parallel", "Basic K-Means", and "DB-Scan". Their proposed method was done on the same data sets to calculate their execution time and accuracy. It enhances performance by reducing the network traffic using partition, aggregation.

Jerril M. et al. [22] design a proposed algorithm of the parallel K-means algorithm based on map reduce on Hadoop. Their paper compared the performance of evaluation criteria called speedup, scale-up, and Size-up. Speedup tries to evaluate the efficiency of the parallelism to improve the execution time. Scale-up checks the ability of it to grow both the Map-Reduce system and the data size, that is, the scalability of the Map-Reduce tool. Size-up estimates the capacity of it to handle growth. It estimates measurements that take to execute the parallel tasks. According to their opinion, the parallel implementation of K-Means gives better results than sequential K-Means algorithm.

3. MAP REDUCE PARADIGM

Map Reduce is the software paradigm for processing larger massive and scalable dataset in the cluster. Map Reduce model processes the unstructured dataset available in a clustering format. Map Reduce is a most popular model used for processing a large set of the data in a parallel and distributed clustering algorithm. It offers numbers of benefits to handle large datasets such as scalability, flexibility and fault tolerance. The map-reduce framework is widely used in processing and managing large data sets. It is also used in such applications

like document clustering, access log analysis, and generating search indexes. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two essential tasks, namely Map and Reduce.

A. Mapper Phase

The Map-Reduce framework is commonly used to analyze enormous datasets like tweets sets, online texts or large scale graphs. The Mapper and Reduce are two essential phases in MapReduce algorithm. Firstly, the mapper phase starts the execution of the map-reduce program. The large dataset that passed into a mapping function to create similar small datasets which called chunk [12]. The Mapper uses a list of key/value pairs, then processes all of it. The mapper produces zero or more (key/value) pairs. The mapper phase output contains the key and the value of the number of instances that lied in the dataset. This structure gives a smooth and stable interface for programmers to resolve large-scale clustering difficulties.

Algorithm 1 : Mapper Function

```
Store samples dataset
Do
  Read mapper-data from samples dataset one by one
  Do
    clustered data=k-means(samples dataset, K,
    distance);
    Send clustered data to the Reducer
  End Of Mapper-Data
  While end-of-file
  Call reducer
End
```

B. k-means Clustering Phase

Clustering is considered a core task of exploratory data analysis and applications of data mining. Clustering task is grouping objects' sets in a way that objects in the same group (a cluster) are similar to one another than to those in other groups (clusters) [9] [11]. The Partition clustering is a widely method where a number of objects are set and the data sets are partitioned into a number of clusters in which each cluster includes similar objects.

The k-means algorithm is used extensively for clustering large datasets. The concept is classifying a presented set of data into k number of disjoint clusters, in which the value of k is fixed in advance. The k-means algorithm [6] is effective for many practical applications in producing clusters. However, the traditional k-means algorithm is extremely high in computational complexity, particularly for large sets of data. Moreover, different types of clusters result from this algorithm depending on the random choice of initial centroids. Many attempts were made by researchers to improve the k-means clustering algorithm performance. This paper proposed a method for improving the accuracy and efficiency of the k-means algorithm. It is used widely due to the ability to produce better cluster results compared to other clustering techniques plus its fast computation.

Algorithm 2 : K-means

Given: dataset of element ($e_1, e_2 \dots e_n$). K : no clusters,

Target :Split the n data elements into k ($\leq n$) partitions $P = \{P_1, P_2, \dots, P_k\}$

1. Set Initial mean value for k cluster randomly.

2. Assign each data element to closest mean.

$$(t) = \left\{ e_p : \|e_p - m_i^{(t)}\|^2 \leq \|e_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k \right\}, p_i$$

3. When data elements have been assigned, the centroid of each of the k clusters becomes the new mean.

$$m_i^{(t+1)} = \frac{1}{|p_i^{(t)}|} \sum_{j \in p_i^{(t)}} e_j$$

4. Repeat Steps 2 and 3 until when the assignments no longer change.

C. Reducer Phase

The reducer phase is the main second part of map reduce. It is responsible for collecting the results coming from mappers. The reducer has three steps; Shuffle, Sort and Reduce.

Shuffle step which receives the output from a mapper phase as input and merges these result tuples into a smaller collection of tuples. In the sort, step values are sorted according to the key. Shuffle and sort process is sent in parallel. The last step here calls the reduce method that takes $\langle \text{key}, \text{list of corresponding value} \rangle$ pair and produces the output into the file system.

The reduce phase created a single output. There are multiple reducers to parallelize the aggregations. Finally, MapReduce is considered easier to scale data processing over various computing nodes.

Algorithm 3: Reducer function

1. Store clustered data

2. Generate cluster label Vector for clustered data

3. Generate Output Matrix $M \times D$, where M is mappers no. & D is clusters no.

4. Initialize Output label Matrix to all cluster.

5. get output from all MAPPERS

While hasNext (intermediateValuesIn)

Put outcome from MAPPERS into output

Matrix(i)= Output

Allocate cluster label according to cluster vote
function Cluster label=Cluster vote(output)

End While

1) Fuzzy Gaussian membership function

The Gaussian fuzzy membership function is considerably famous in the fuzzy logic literature. It considered the main connection between the fuzzy systems and the radial basis function (RBF) of neural networks. Also, the Gaussian is used to represent vague, linguistic terms. It focuses on an adaptive distance measure; it can adapt the distance norm to the

underlying distribution of the data which is presented in the different sizes of the clusters [1]. Gaussian functions are exercised in statistics to describe the standard distributions. It used in signal processing to represent Gaussian filters. In image processing where two-dimensional Gaussians are performed for Gaussian blurs, in mathematics to solve equations and diffusion equations to define the Weierstrass transform.

Algorithm 4 :Fuzzy Gaussian Membership

1. get label Cluster matrix from all mappers

2. Generate a matrix $M \times N$ contains cluster label

M is cluster no. N is number of mappers

3. Do

1. Compute "mean" & "standard deviation" for every clustered data

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

2. Compute "Membership function" for every cluster

$$\mu_i(x) = e^{-\frac{(x-\mu)}{\sigma}}$$

$i = 1, 2, \dots, M$, and M is number of clusters

3. allocate the cluster label with the greatest membership function

$$\text{cluster} = \text{Max}(\mu_i(x))$$

Until End Of File

2) Mode function

It is the majority vote, the concept of mode makes sense for any random variable estimating values from a vector space, containing the real numbers and the integers. The mode-function is quickly comprehensible and accessible to calculate. The clustered label is allocated according to the majority of the clustered data.

Cluster label = mode (output)

$$\text{mode} = \arg \max[\text{Output}]$$

4. PROPOSED MODEL

K-means algorithm is based on determining an initial number of iterations, and iteratively reallocates objects among groups to convergence. The proposed model based on k-means and handled by map-reduce programming model.

The proposed model in this paper consists of two phases as shown in Fig.1 that namely, Mapper Phase and Reduce Phase. The first phase split the big dataset into small groups which called mapper according to RAM capacity. Next, the significant part had started when K-Means received the data from the mapper and return cluster label.

The second phase called reducer phase. In this phase used the Fuzzy Gaussian algorithm and Mode function. It had started after receiving cluster label. So, it collects them to produce one output.

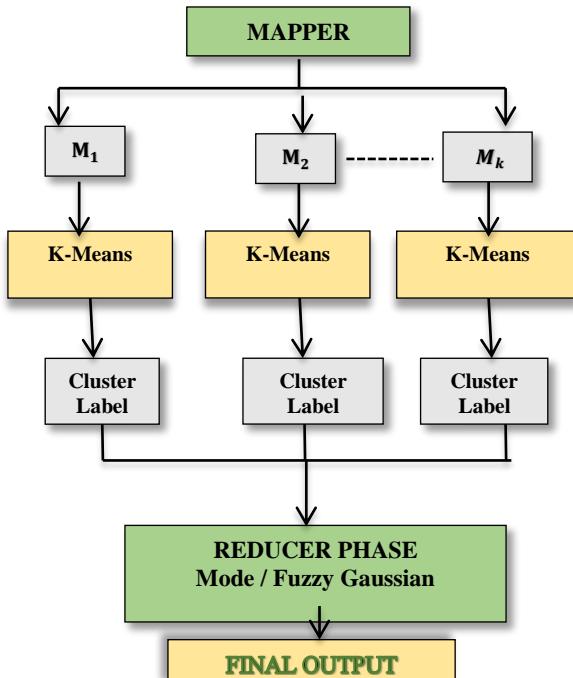


Fig 1 : The flow chart of proposed model.

5. EXPERIMENTAL RESULTS

Data mining algorithms have two important performance indicators are the accuracy for cluster data and the time taken to apply the training.

The propose approach had developed map reduce tool to implement the approach. The experiments are performed on a machine having Intel core i7 processor with 16 GB RAM and windows 10 OS. MATLAB R2014b is used in the experiments.

A. Dataset

In this paper, the datasets had taken from [10]. It is four openly available datasets; Table1 shows the main features of these datasets [10].

TABLE I. DATA SET DETAILS

DatasetName	Records-no.	Attributes-no.	Classes-no.
"Covtype"	581012	54	7
"Covtype-2"	581012	54	2
"Poker Hand"	1025009	10	9
"Poker Hand-2"	1025009	10	2

The Covtype dataset contains 581012 sample to predict forest cover type from cartographic variables. Any individual relates to one of seven categories (classes) such as "Spruce/Fir, Lodgepole Pine, Ponderosa Pine, and Cottonwood/Willow." The second one is "Covtype-2". It is similar to Covtype except for the number of class (2 class).

Each instance of the Poker-Hand dataset is an illustration of a hand containing five playing cards that drawn from a standard deck of 52. Suit and Rank are two attributes which represent every card, for a total of 10 predictive characteristics. The order of cards is essential, which is why there are 480 possible Royal Flush hands rather than 4. Also, the "Poker Hand-2" is similar Poker Hand except the number of classes is two classes.

B. Results

In this part, experiment's results that are obtained after the implementation of K-means in mapper phase and using two different functions in reducer phase. The four different datasets had applied in the experiments.

TABLE II. ACCURACY AND TIME TAKEN BY TRAINING THE K MEAN AND MODE ALGORITHM

Method \ Data sets	Covtype	Covtype-2	Poker	Poker-2
Accuracy (%)	56.72	62.1	62.67	63.2
Time Taken (Ratio)	7.20126	6.725463	6.012545	6.21541

The results obtained by using Mode function is shown in table 2. The proposed approach achieves 56.72% accuracy in time 7.20126 in case of using "Covtype" dataset which is considered the lowest accuracy and highest time taken. The accuracy improved by 5.38% when decrease the number of classes using "Covtype-2".

TABLE III. ACCURACY AND TIME TAKEN BY TRAINING THE KM& FUZZY GAUSSIAN ALGORITHM

Method \ Data sets	Covtype	Covtype-2	Poker	Poker-2
Accuracy (%)	62.1	75.6	63.4	73.4
Time Taken (Ratio)	8.01245	8.12542	7.124512	7.124512

The results obtained by using Fuzzy Gaussian are shown in table 3. The best accuracy is 75.6% using "Covtype-2" which enhance results achieved using "Covtype".

Figure 2 shows the comparison between the mode and fuzzy function accuracies have been utilized in reducer phase. The results show improving using fuzzy Gaussian than mod function by leading to simple and straightforward linear algebra implementations. In case if using all " Covtype"," Covtype-2"," Poker "," Poker-2" respectively . The accuracy results indicate that Fuzzy Gaussian is better that Mode function this probably because of allowing one to quantify uncertainty in predictions resulting not just from intrinsic noise in the problem but also the errors in the parameter estimation procedure.

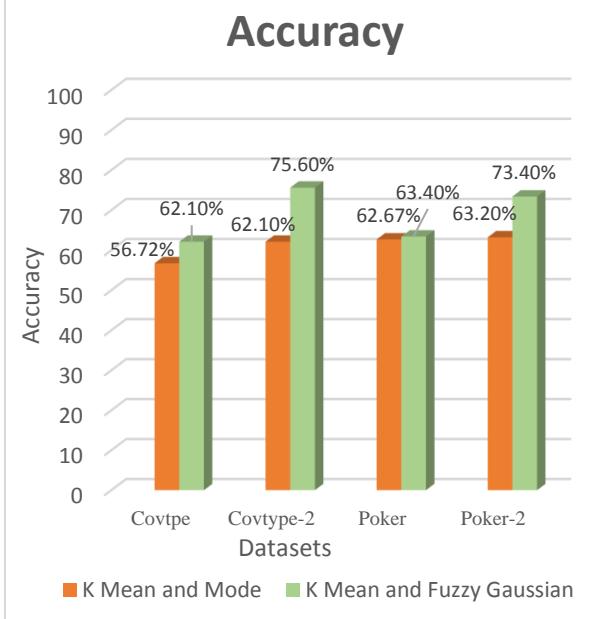


Fig. 2: Accuracy Comparsion for four data set

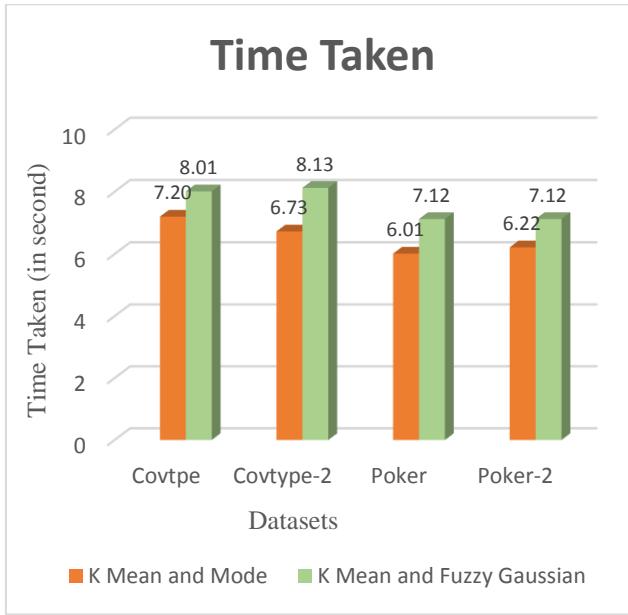


Fig.3: Run Time Comparsion for four data set

Time taken comparisons between the mode and fuzzy function is shown in Figure 3. The results show that the time taken by fuzzy Gaussian is higher than mod function time this probably because of all calculation takes too much time to calculate it more than Mode function.

6. CONCLUSION

The proposed approach is based on the MapReduce programming model. It consist of two phases, Mapper and Reduce phases. In Mapper phase; it had distributed to a mappers group using the map function. K-means is applied on

small datasets which existed in mappers. In Reduce phase; the reduce function is resulted by combining outputs using "Mod" and "Fuzzy Gaussian" functions. Gaussian function includes mixed membership; each cluster can have unconstrained covariance structure. Think of rotated or elongated distribution of points in a group. The cluster assignment is flexible. All instance belongs to each cluster to a different degree. The degree is according to the probability of the instance which generated from each cluster's (multivariate) normal distribution. Experimental results showed that the proposed approach gives higher accuracy when using "Fuzzy Gaussian" function than using "Mod" function, as well as perfect time was taken. Also, Fuzzy Gaussian proved its efficiency in accuracy than Mod but with more time in execution.

REFERENCES

- [1] Agnes Vathy-Fogarassy, Attila Kiss and Janos Abonyi, "Hybrid Minimal Spanning tree based clustering and mixture of Gaussians based clustering algorithm", pp. 313-330, Springer, 2006.
- [2] Neha D., B.M. Vidyavathi, "A Survey on Applications of Data Mining using Clustering Techniques", International Journal of Computer Applications, vol.126, no.2, 2015.
- [3] Anshul Yadav, Sakshi Dhingra, "A Review on K-means Clustering Technique", International Journal of Latest Research in Science and Technology, vol.5, Issue 4, no.13-16, 2016.
- [4] Vinod S. Bawane, Deepti P. Theng, "Enhancing Map-Reduce Mechanism for Big Data with Density-Based Clustering", International Journal of Innovative Research in Computer and Communication Engineering, vol.3, Issue 4, 2015.
- [5] Kosha Kothari, Ompryka Kale, "Survey of Various Clustering Techniques for Big Data in Data Mining", IJIRT, vol.1, Issue 7, 2014.
- [6] Y. S. Thakare, S. B. Bagal, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics", International Journal of Computer Applications, vol.110, no. 11, January 2015.
- [7] SK Ahammad Fahad, Md. Mahbub Alam, "A Modified K-Means Algorithm for Big Data Clustering", IJCSET, vol.6, Issue 4, 129-132, April 2016.
- [8] Chowdam Sreedhar, Nagulapally Kasiviswanath, Pakanti Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop", Journal of Big Data 4, no. 1, 2017.
- [9] T. Sajana, C. M. Sheela Rani and K. V. Narayana, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, vol.9, no.3, 2016.
- [10] the UCI repository
- [11] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering vol.1, 2009.
- [12] Sergio Ramírez-Gallego, Alberto Fernández, Salvador García, Min Chen, Francisco Herrera, "Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce", Information Fusion, vol. 42, pp.51-61, 2018.
- [13] Mohammed S. Hadi, Ahmed Q. Lawey, Taisir E. H. El-Gorashi and Jaafar M. H. Elmirghani, "Big Data Analytics for Wireless and Wired Network Design: A Survey", 2018.

- [14] Anju, Preeti Gulia, "Clustering in Big Data: A Review", International Journal of Computer Applications, vol.153, no.3, pp.44-47, 2016.
- [15] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S. and Bouras, A., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis", IEEE transactions on emerging topics in computing, vol.2, no.3, pp.267-279, 2014.
- [16] Yangyang Li, Guoli Yang, Haiyang He, Licheng Jiao, Ronghua Shang, " A study of large-scale data clustering based on fuzzy clustering", Soft Computing, vol.20, no.8, pp.3231-3242, 2016.
- [17] Srikanta Kolay, Kumar S. Ray, Abhoy Chand Mondal , "K+ Means : An Enhancement Over K-Means Clustering Algorithm", arXiv preprint arXiv:1706.02949, 2017.
- [18] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng, "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing, no.1, pp.67, 2016.
- [19] Anju Abraham, Shyma Kareem," Security and Clustering Of Big Data in Map Reduce Framework:A Survey", International Journal of Advance Research, Ideas and Innovations in Technology, vol.4, Issue 1, 2018.
- [20] Mugdha Jain, Chakradhar Verma, " Adapting k-means for Clustering in Big Data", International Journal of Computer Applications, vol. 101, no.1, 2014.
- [21] G. Venkatesh, K. Arunesh, "Map Reduce for big data processing based on traffic aware partition and aggregation", Springer Science and Business Media, 2018.
- [22] JERRIL MATHSON MATHEW, JYOTHS JOSEPH " Parallel Implementation of Clustering Algorithms Using Hadoop", International Journal of Advances in Electronics and Computer Science, vol.3, Issue 6, 2016.

Cross Layer based Hybrid fuzzy ad-hoc rate based Congestion Control (CLHCC) approach for VoMAN to improve quality of VoIP flows

V.Savithri¹,¹ Assistant Professor, Part-Time Ph.D. Scholar, Bharathiar University, Coimbatore, India.

Dr.A.Marimuthu², Associate Professor& Head, PG and Research Department of Computer Science, Government Arts College(Autonomous),Coimbatore, India.

Abstract - Mobile environment pretense a number of novel theoretical and optimization issues such as position, operation and following in that a lot of requests rely on them for desirable information. The precedent works are sprinkled across the entire network layer: from the medium of physical to link layer to routing and then application layer. In this invention, we present outline solutions in Medium Access Control (MAC), data distribution, coverage resolve issues under mobile ad-hoc network environment based on congestion control technique using Transmission Control Protocol (TCP). In mobile ad-hoc network issues can arise such as link disconnections, channel contention and recurrent path loss. To resolve this issue, we propose a Cross Layer based Hybrid fuzzy ad-hoc rate based Congestion Control (CLHCC) approach to maximize network performance. Based on the destination report it regulates the speed of data flow to control data loss by monitoring the present network status and transmits this report to the source as advice. The source adjusts the sending flow rate as per the advice. This is monitored by channel usage, ultimate delay, short term throughput.

Index Terms— MAC, MPSD, RTT, SIFS TCP,

I. INTRODUCTION

Mobile Ad-hoc Network (MANET) is a transient dynamic network formed by set of mobile nodes. Routing in MANET is a complex and challenging task because of its dynamic nature, link stability and infrastructure less concept. A packet might traverse many intermediate nodes supporting dynamic link to reach the desired destination. Routing algorithm should generate feasible route by collecting routing status and route packets over the optimal route to support Quality of Service (QoS).

Voice over MANET is becoming important day by day as users demand to use real time applications. As Quality of Service (QoS) issues in infrastructure based networks still remain unsolved, it is a challenging task in MANET that needs to be solved for real time applications [12]. Increase in real time traffic increases network congestion. Congestion control policies are classified into three types namely window based, rate based and hybrid. Window based congestion control policy adjusts the congestion window as per the changing network status. Rate based congestion control policy increases or decreases the data rate of the sender as per the current status of the network [8]. Hybrid approach combines both the above discussed policies to control congestion.

This paper is organized as follows: Section II describes the literature survey and Section III presents the methodology of Fuzzy based Congestion Control approach using voice over TCP. Section IV presents the result analysis of the new approach. Section V summarizes the most important simulation results and their interpretation. Finally, Section VI concludes this paper.

II. LITERATURE SURVEY

Security in mobile ad hoc network is hard to accomplish due to vibrantly changing and fully decentralized topology as well as the vulnerabilities and limitations of wireless data transmissions[15].A lot of research has since focused on mechanisms to improve TCP performance in cellular wireless systems[1]. In MANET, due to dynamic nature of the network and variable number hops between source and destination, the fairness and efficiency get degraded. It is also seen that the existing protocols are not good enough in setting its parameters like congestion window, congestion window limit, round trip time and the retransmission timeout timers [11]. The routing overhead of such an algorithm increases with the square of the number of mobile nodes in a MANET[2]. A hybrid routing algorithm that combines the merits of existing protocols that can be used to address this issue of growth in network size and load balancing whose behavior can be modified according to the size of network [9]. The effect of high bit error rate and route re-computation on the performance of TCP in mobile ad hoc network is analyzed [4]. In particular, it allows the routing protocol to operate more efficiently by reducing the control traffic in the network and simplifying the data routing [7]. In order to avoid congestion in the network, it is required to use an efficient congestion control algorithm for successful transmission of data throughout the network [13]. Numerous examinations have been done to enhance the

execution of TCP at network layer by enhancing the routing strategy. M-ADTCP is another method that presents a Modified AD-hoc Transmission Control Protocol where the receiver detects the probable current network status and transmits that information to the sender as feedback. The sender behaviour is altered appropriately [6]. In this way, there is an extension to enhance the execution of TCP at transport layer [14]. TCP supplies end-to-end reliable delivery of data between an application process on one computer and the process running on another computer, by adding services on top of IP [5]. In order to use these limited resources efficiently, an intelligent routing strategy is required which should also be adaptable to the changing conditions of the network, like, size of the network, traffic density and network partitioning[10].

III. A. CLHCC METHODOLOGY

Application Layer
VoIP based Communication
Transport Layer
Congestion detection and Controller System
Network Layer
Computation of Middle Packet Setback Distinction, Short Term Throughput, Delay
Physical and MAC Layer
Computation of Channel business ratio

Figure 1. Cross Layer based VoMAN Architecture

We consider a mobile ad-hoc network with n nodes S (s1; s2; . . . sn). These n nodes are randomly disseminated in a field. At random time periods, each node

can send event values (E_V) from the field at its location and sends them E_V to the destination. Nodes send a network wide broadcast message at every particular interval, to all nearby nodes in the network. Nodes built with a global positioning system (GPS) recipient at an exactly known position which is used to receive collected information from nearby nodes. The nodes are installed randomly and they may organize by them self as hop based jointly to complete a transmission sensed the environment. Detection of nearby nodes is accomplished by a periodic HELLO messages sharing among the nodes, produced at fixed interval. When a node receives HELLO message it updates the neighbor table NT.

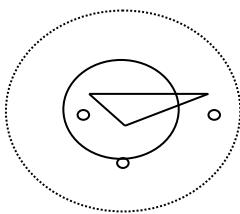


Figure 2. Transmission Area

If node does not receive HELLO message within the fixed time interval at each sequence purge it from NT. By this way update nearby routing node freshly. Mobile devices are deployed in the network region randomly and moves anywhere in the network using the random way point mobility model in non uniform manner. Source node initiates the data transmission by sending the first control packet to the destination.

In network layer, device looks for the routing table (RT) entry to forward the packet to the destination. If there is no route found in RT, then the device executes the route discovery process to identify the route to connect with the destination. When the source device wants to generate a new path to the final destination, the sender broadcast a route request packet to the entire network and it floods in

many directions by its neighbors obtained from NT. When the neighbors collect the RREQ it produces a reverse path to the source device. By these way neighbor devices of next hop to the sender established. While forwarding the RREQ and RREP the reverse entry and forwarding entry is formed respectively. The middle devices re-broadcast the RREQ to their neighbors and update the device address as value[i] in their RT as

```
RTTable.count++
ΣRT += RTTable.value[i]
```

These message exchanges will be used to form the reverse path for route reply RREP from the destination. The devices collecting these packets are cached from source and when the link is disconnected by using this it sends RERR packet which holds information about devices that are unable to access. After constructing the path, the devices forward the data according to the constructed path from source to destination.

Due to mobility, if link failure occurs then the route reconstruction process is invoked to identify the alternate route to connect to the destination device. During data transmission, a packet either gets dropped or gets delayed due to network congestion.

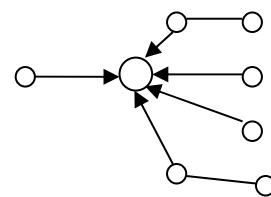


Figure 3. Router Congested Node

In transport layer, congestion control mechanism is operated to identify and to control the network congestion in base model using the dedicated parameters such as

Middle Packet Setback Distinction (MPSD) and temporary bits per second (TB). In this Initial MPSD = 0.

Update MPSD count++

The Sum of MPSD is computed using the formula

$$\Sigma MPSD += MPSDValue[i];$$

Average of MPSD is computed using the formula

$$AvgMPSD = \frac{\Sigma MPSD}{MPSDCount}$$

$$D\Sigma MPSDCount++$$

$$D\Sigma MPSD += (AvgMPSD - MPSDValue[i])^2$$

Computed variance of MPSD

$$VarianceMPSD = D\Sigma MPSDCount / MPSDCount$$

For MPSD an upper bound (UNMPSD) and lower bound (LBMPSD) are defined as threshold points.

$$StddevMPSD = \sqrt{VarianceMPSD}$$

$$LBMPSD = AvgMPSD - StddevMPSD$$

$$UBMPSD = AvgMPSD + StddevMPSD$$

If the current statistics of the network crossed these threshold boundaries then the network is marked as congested network.

$$if(MPSD < LBMPSD || MPSD > UBMPSD)$$

Enable congestion Flag = 1

Also the TB computed as

$$TBCount++$$

$$\Sigma TB += TBValue[i];$$

$$AvgTB = \Sigma TB / TBCount;$$

$$D\Sigma TB += (AvgTB - TBValue[i])^2$$

MBS is computed based on idle packet state; success packet state and collision state at channel usage period and are also based on the number of transmissions as success, collision and idle states as given below. Channel usage (CU) estimation provides the present utilization of overall bandwidth consumption.

$$PI = pow((1-tow),n);$$

$$PS = N - (1 - NI)^2$$

$$PC = 1 - Pidle - Psuccess;$$

$$TS = TData + TAck + SIFS + DIFS$$

$$TC = TData + Timeout + DIFS$$

$$TI = 1 - TSuc - Tcol$$

$$CU = \frac{PSxTS}{(PIxTI+PSxTS+PCxTC)}$$

$$MBS = \frac{(PSxTS + PCxTC)}{(PIxTI+PSxTS+PCxTC)}$$

Due to link failure and node failure, these parameters may provide false deviations as the congested state of the network. In order to avoid these false positive cases, the metrics are extended as follows, Completion Time (CT) is taken as

$$D\Sigma CT += (AvgCT - CTValue[i])^2,$$

Packet Loss Case (PLC) is computed as

$D\Sigma PLC += (AvgPLC - PLCValue[i])^2$, and Packet Delay (PD) along with the Medium Busy State(MBS).

$$PD = \frac{\Sigma PD}{PDCount}$$

For these extended parameters, both lower and upper bound are defined as threshold limits which are used to classify the current network situation as congested or not. If it is congested then congestion flag is enabled.

Update MPSDadd(MPSD)

Update TBAdd(TB)

Enable CongestionFlag

$$CHMBSTime = b * AavgMBS + (1-b) *$$

CurrentMBSTime

$$CHDS = b * AvgDelay + (1-b) * CurrentDelay$$

CH NewRate = NewRate

Each parameter is computed using the statistical model related to the Gaussian distribution with mean and standard deviation. To compute these parameters, the corresponding information is maintained in the newly added header called TCP Congestion header (CH). CH is computed as the time difference between data packet sent and acknowledgment received.

$$AvgDelay = \alpha AvgDelay + (1-\alpha) x CHDS$$

$$CHMBS Time = Max(CHMBSTime, MBS)$$

$$CurrenMBSTime = CH MBSTime$$

$$CurrentDelay = CHDS$$

Source node attaches the send time information in CH, destination receives the information and copy it in CH of ACK message. Once the ACK message reaches the source end, it calculates the difference between the ACK packet received time and Data Packet sent time as CH. MPSD is computed as average delay. Transmission delay is computed as the time difference between the packets sent and packets received for both data and acknowledgment messages.

$$AvgDelay = \alpha * AvgDelay + (1-\alpha) * CHDS$$

$$CHMBS Time = Max(CHMBSTime, MBS)$$

$$CurrenMBSTime = CH MBSTime$$

$$CurrentDelay = CHDS$$

If (Enclosed Reaction == 1)

PLC is computed using the packet sequence number difference at the destination end. It is calculated as the ratio between the number of packets not in order and the total number of packet transmitted. TB is computed as average number of bits transmitted in cycle duration.

Using the current data rate, data packet size and acknowledgment size, approximate channel occupancy time are computed for both data packet and acknowledgment packet. Back-off timeout period is also computed using the channel occupancy time, propagation delay and retransmission delay. Using the individual slot time for each node, idle probability, transmission successful probability and collision probability are estimated. The successful transmission duration is computed as the sum of Data duration, Acknowledgment duration, SIFS and DIFS periods.

The collision duration is computed as the sum of data duration, timeout period and DIFS periods. From the unit slot duration, idle slot period is computed using the difference between unit with sum of successful transmission duration and collision duration. MBS is computed using the ratio product of successful transmission duration and successful transmission probability with entire probability duration. MBS is computed as the ratio product of successful transmission duration with collision duration

and successful transmission probability along with collision probability to the entire probability duration as delay.

```
If(PacketType == ACK)
If(CongFlag == 1)
Compute QueueDelay = CURRENT_TIME - CHTime
CHΣDs += QueueDelay
CHCountDs++
CHDS = CHΣDs / CHCountDs
If(AveDelay == 0)
AvgDelay = CHDS
```

The MBS is estimated in Mac layer and attached in the CH header field named MBS metric.

B. Fuzzy Inference System

The fuzzy inference system operates at destination node. After collecting delay, ongoing rate and MBS parameters, they are given as an input to the FIS based congestion detection system

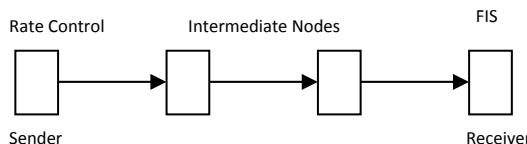


Fig.4 Fuzzy based Congestion Detection System

which employs fuzzification, rule set matching and detection of congestion state. In fuzzification stage, using the lower and upper bound the input parameters are converted into fuzzy linguistic variables.

```
FuzzyInput = MBS Delay RS
FuzzyRule = RuleSetMatching (Fuzzy1,Fuzzy2,Fuzzy3)
CongState = CongestionState(FuzzyRule)
Set CongState
```

The linear relationship growth model is used in rule set formation since these parameters have uniform deviations in the parametric values. Identified fuzzy linguistics are compared with the rule set to compute the exact match in the rule set to select current state of congestion. Congestion states are classified as VLOW, LOW, MEDIUM, HIGH, VHIGH.

```
CongestionState (FuzzyRule)
if(rule == VLOW)(0,0.2)
if(rule == LOW) (0.2,0.4)
if(rule == MEDIUM) (0.4,0.6)
if(rule == HI)(0.6,0.8)
if(rule == VHI)(0.8,1.0)
```

If congestion is detected in the network state, then the CH is marked in the acknowledgment message. Once the acknowledgment packet with congestion notification reaches the source node, it reduces the current data rate to decrease the outgoing packet count.

If congestion state is VLOW or LOW then destination calculates new data rate using multiplicative increase. If congestion state is MEDIUM then destination calculates new data rate using additive increase. If congestion state is HI or VHI then it calculates new data rate using multiplicative decrease. It updates the new data rate and sends the acknowledgement packet to sender. Intermediate nodes increase or decrease congestion window size as per the new data rate they receive from the neighbor.

C. E-MODEL

E-model is the common ITU-T transmission rating model. This computational model can be used to measure the quality of Voice data that help ensure that users will be satisfied with end-to-end transmission performance. The

main output of the model is a scalar rating of transmission quality.

It is the most commonly used method for predicting the quality of voice signal on user side.

The ITU-TG.107 defines the relationship between R and MOS as follows:

$$\text{MOS} = 1 \text{ for } R < 0$$

$$\text{MOS} = 1 + 0.035R + R(R-60)*(100-R) / 7 * 0.000001 \text{ for } 0 < R < 100$$

$$\text{MOS} = 4.5 \text{ for } R \geq 100$$

The rating factor R is composed of the basic formula:

$$R = Ro - Is - Id - Ie\text{-eff} + A$$

Ro (signal to noise ratio) is a mathematical summary of how the voice levels compare to the different noise sources including circuit noise and room noise.

Id (delay impairments) is a mathematical summary of transmission delay, talker echo and sidetone.

Is (simultaneous impairments) considers non-optimum sidetone, quantizing distortion, overall loudness and other impairments which occur more or less simultaneously with the voice transmission.

Ie (equipment impairment) and A (Advantage Factor) are both single value quantities.

To assist with calculations, default values and permitted ranges have been established.

This calculates MOS value based on the given criteria[3].

The following table shows the relationship between R factor and user satisfaction.

R Factor	MOS	User Satisfaction
140	4.5	Very Satisfied
80	4.024	Satisfied

120	4.5	Very Satisfied
138	4.5	Very Satisfied
110	4.5	Very Satisfied

IV. RESULT ANALYSIS

Simulation compares the proposed approach CLHCC with FARCC and ADTCP by varying the number of nodes and simulation time. Results show that the performance of the proposed approach at network level and at user level. It outperforms well than the two approaches in terms of packet delivery ratio, end-end-delay, through put. It calculates user level quality called Quality of Experience (QoE) using E-Model. MOS values show that the proposed approach attains user dissatisfaction level in terms of quality.

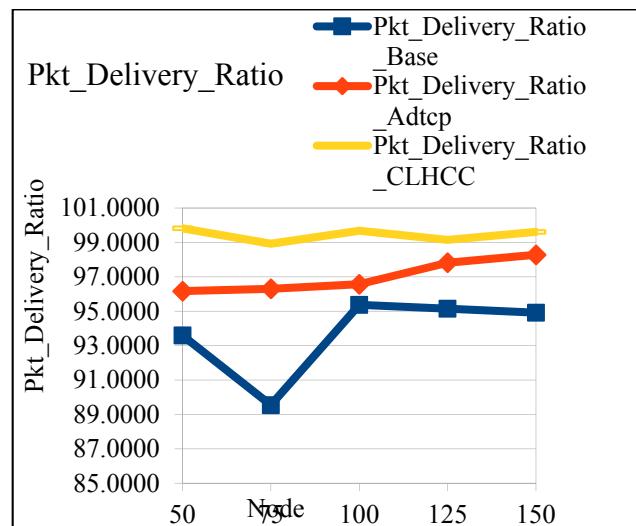


Fig 5 : Node Versus Packet Delivery Ratio

Figure 5 shows the comparison between three approaches based on the metric Packet Delivery Ratio. The graph clearly indicates that the proposed approach outperforms well than the existing approaches.

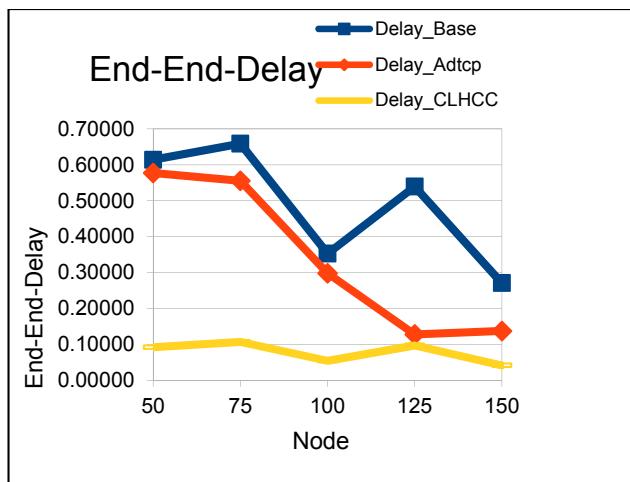


Fig 6 : Node Versus End-End-Delay

Figure 6 shows the performance of proposed approach in terms of End-End-Delay. It suffers from less End-End-Delay than the existing approaches.

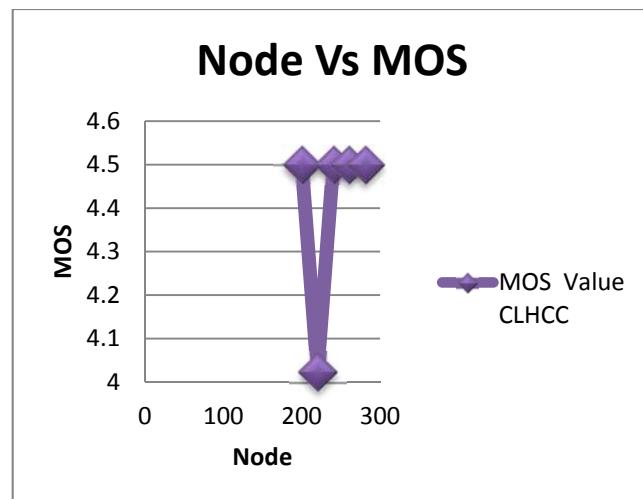


Fig: Node Versus MOS

Figure 7 shows that user level quality of Voice. Quality remains at same level as the performance increases.

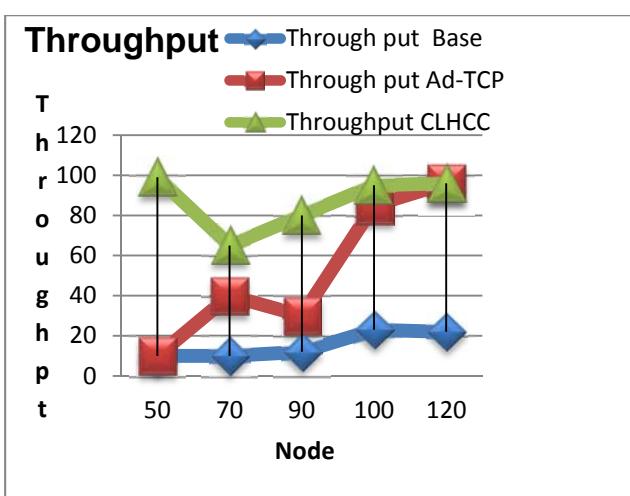


Fig 7 : Node Vs Throughput

Figure 7 shows the maximum throughput attainment of proposed approach.

V. CONCLUSION

In Mobile Ad-hoc Network, MBS, TB, Delay distinction are used to measure the congestion level of the network. These investigations support us to change transmission rate. In this paper, we checked congestion level in MANET with Fuzzy Inference System to finalize the data rate and to control the flow of data at sender by the receiver dynamically. It also adjusts the congestion window as per the new data rate on the stamp field. Hence it supports both proactive and reactive approach to congestion detection and to control congestion. The proposed approach outperforms well for network level quality and user level quality than the existing congestion control approaches ADTCP and FARCC.

REFERENCES

- [1] Gavin Holland and Nitin Vaidya, "Analysis of TCP Performance over Mobile AdHocNetworks", ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom '99), (Seattle, Washington), August 1999.

- [2] Jane y. yu and Peter h. j. Chong, "A Survey of Clustering Schemes for Mobile ad hoc Networks", IEEE Communications Surveys & Tutorials, First Quarter, 7, 2005.
- [3] Lingfen Sun, Emmanuel C. Ifeachor, "Voice Quality Prediction Models and Their Application in VoIP Networks", IEEE Transactions On Multimedia, Vol. 8, No. 4, August 2006.
- [4] Foez ahmed, Sateesh Kumar Pradhan, Nayema Islam, and Sumon Kumar Debnath, "Performance Evaluation of TCP over Mobile Ad-hoc Networks", International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010.
- [5] Praveen Dalal, "Study on Transport Layer Protocols for Wireless Ad-Hoc Network", Proceedings of the 5th National Conference; INDIACom Computing For Nation Development, March 10, 2011.
- [6] Sreenivasa B.C, G.C. Bhanu Prakash, K.V. Ramakrishnan, "Comparative analysis of ADTCP and MADTCP: Congestion Control Techniques for improving TCP performance over Ad-hoc Networks", International Journal of Mobile Network Communications & Telematics (IJMNCT) Vol.2, No.4, August 2012.
- [7] Abdelhak Bentaleb, Abdelhak Boubetra, Saad Harous, "Survey o Clustering Schemes in Mobile Ad hoc Networks", Scientific Research Communications and Network, May 2013.
- [8] H. Zare, F. Adibnia, V. Derhami, "A Rate based Congestion Control Mechanism using Fuzzy Controller in MANETs", International Journal of Computer Communication, June 2013.
- [9] Gargi Parashar, Manisha Sharma, "Congestion Control in Manets Using Hybrid Routing Protocol", IOSR Journal of Electronics and Communication Engineering, Vol.6, Issue 3, Jun 2013.
- [10] Bandana Bhatia, Neha Sood, "AODV based Congestion Control Protocols: Review", International Journal of Computer Science and Information Technologies, Vol.5(3), 2014.
- [11] Waleed S. Alnumay, "Security Enhanced Adaptive TCP for Wireless Ad Hoc Networks", Journal of Information Security, 2014.
- [12] Said El brak, Mohamed El brak, Driss Benhaddou, "A New QoS Management Scheme for VoIP Application over Wireless Ad Hoc Networks", Journal of Computer Networks and Communications, Vol. 2014.
- [13] Abinasha Mohan Borah, Bobby Sharma, Manab Mohan Borah, "A Congestion Control Algorithm for Mobility Model in Mobile Ad-hoc Networks", International Journal of Computer Applications, Volume 118 – No.23, May 2015.
- [14] Kaushika Patel, Nayana Ram, Virendra Barot, "TCP in MANET : Challenges and Solutions", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 12, December 2015.
- [15] Ritika Mehra and Manjula Saluja, "Adaptive Congestion Control Mechanisms in Mobile Ad-Hoc Networks", International Journal of Engineering Development and Research, Volume 5, Issue 1, 2017.

Caesar Cipher Method Design and Implementation Based on Java, C++, and Python Languages

Ismail M. Keshta 
ismailk@dcc.kfupm.edu.sa
dr.ismail.keshta@gmail.com

Abstract—Today information security is a challenging factor that touches a lot of areas, including computers and communications. Message communication is kept secure through cryptography so that an eavesdropper is not able to decipher a transmitted message. One of the oldest and simplest known algorithms for cryptography is the Caesar cipher algorithm. In this paper, three programs based on Java, C++, and Python languages have been developed to implement the Caesar cipher algorithm to aid information security students and help them understand this fundamental algorithm. A code flow chart is used for each program to describe the code's flow. It also reveals the sequence of steps for the code's main methods, as well as the relationships between them. Furthermore, various technical descriptions are presented in detail for each of the methods used in both the encoding and the decoding of the messages.

Keywords: *cryptography, Caesar cipher, encryption, decryption, plain text, cipher text*

I. INTRODUCTION

Desiring to securely transmit messages that cannot be understood by others is not new. In fact, people have needed to keep their communications private and secure for centuries [1]. Today digital communication systems are able to carry huge amounts of sensitive data, particularly those related to the Internet—for example, sending information about a credit card in an e-commerce transaction or using e-mail to exchange confidential trade secrets [1][2][3].

Network security is, therefore, a vital aspect of information sharing. Many technological implementations and security policies have been developed in various attempts to remove insecurities over the Internet [4][5]. The total amount of data transferred is not an important factor. What is important is how much security the channel can provide while it is transmitting data [6]. Cryptography is a technique that allows the secure transmission of data without loss of confidentiality or integrity [6][7].

The field of cryptography is vital as it deals with techniques that can convey information securely. Its aim is to allow recipients to receive their message properly while also stopping any eavesdroppers from deciphering and understanding what is written [8]. The original form of the message is called plaintext. The transmitter uses a secure system to encrypt the plaintext to hide the true meaning. This is a reversible mathematical process that produces an encrypted output, which is called ciphertext, and the algorithm used to encrypt the message is called a cipher. The science of breaking ciphers is called cryptanalysis, and a

cryptanalyst tries to break the security of cryptographic systems [8][9].

It is worth noting that a ciphertext can be openly transmitted across the communication channel. However, because of the encrypted nature of this message, eavesdroppers who have access to the ciphertext should not be able to uncover its meaning, as only the intended recipient can decrypt it to recover the original plaintext for interpretation [10]. All these processes are shown in Figure 1.

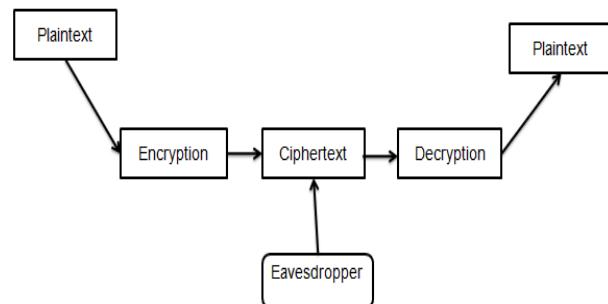


Figure 1: Block diagram of a cryptographic system

Cryptography can be split into two main types based on key distribution: symmetric key cryptography and asymmetric key cryptography. In symmetric key cryptography, the same key is used for both encryption and decryption (see Figure 2). On the other hand, asymmetric key cryptography uses different keys for encryption and decryption (see Figure 3). The two keys are related to each other mathematically, and obtaining one from the other is very hard [11].

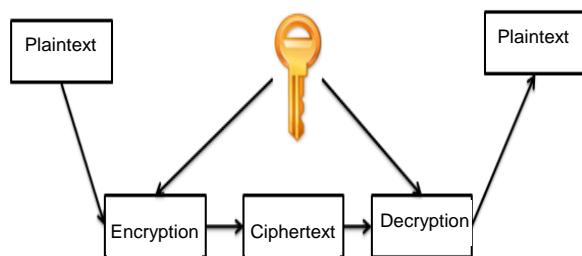


Figure 2: Symmetric key encryption

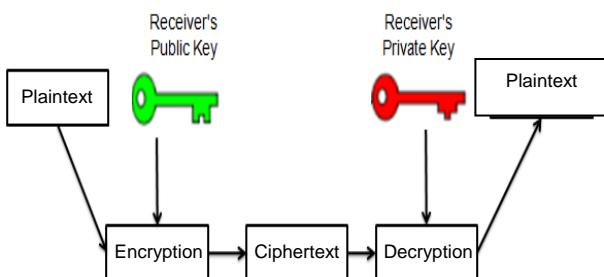


Figure 3: Asymmetric key encryption

Cryptography is a collection of various techniques called “ciphers” that are used to make things more difficult to both read and understand [11][12]. Cryptography has many uses, including Internet security and spying! There is a constant struggle between those who create ciphers and those who set out to break them.

The Caesar cipher is one of the oldest and simplest ciphers. It is no longer used in encryption systems, but strong modern ciphers use it in combination with various other operations [13][14][15]. This results in an algorithm that is more secure than its components. We describe this cipher in further detail in the following section.

II. CAESAR CIPHER

The Caesar cipher is a simple substitution cipher of historical interest. It gets its name from Julius Caesar, who used this code to send his secret messages [16][17]. Caesar encrypted his messages by moving every letter in the plaintext three positions to the right in the alphabet [18]. This cipher is based on shifted alphabets. Therefore, it is also known as a shift cipher, Caesar shift, or Caesar cipher and is used as a substitution method to evolve the encrypted text [19]. The following illustrates Caesar’s method:

Plaintext: the exam will start at eight in the morning
 Ciphertext: wkh hadp zloo vwduw dw hljkw lq wkh pruqlqj

The Caesar cipher encryption process is performed by using the following function [20] [21]:

$$E(x) = (x + n) \text{ mode } 26$$

The decryption process reverses the encryption process by using the following function [22]:

$$D(x) = (x - n) \text{ mode } 26$$

It is worth mentioning that the Caesar cipher method did not last long because of its simplicity and obvious lack of communication security [13]. Breaking this cipher is not difficult as the plaintext’s own statistical information is already contained in the ciphertext. By performing frequency analysis, it is easy to reveal that the Caesar cipher has been used [1] as well as to quickly uncover the message. The amount of the shift $K = 3$ is defined to be the key for the Caesar cipher. Shifts by other amounts can also be used. However, there are just 25 possible shifts in the alphabet, and all possible combinations can be made

quickly [1][13][20]. The fact that a shift in letters is used means that it is not difficult to guess the key because it is a single private key with a space for 1 to 25 different combinations. A long cipher text can be difficult to break manually, and you could have no clue what key has been used, but it has no standing in the modern age of computers anymore as it can be easily broken through a brute force attack because only 25 possible key options available (from $K=1$ to $K=25$) [14].

III. IMPLEMENTATION OF THE CAESAR CIPHER METHOD

Three programs based on Java, C++, and Python languages have been developed to implement the Caesar cipher method using the user-defined key for the shift. Each program starts by asking the user for the source file, the target file, and the key. In particular, it asks the user to enter the file’s name or the path of the file user that is needed for decoding after the decode file path/name is obtained. It then asks for the file name or path required to store the decoding of the plaintext. After both resource files are obtained, the program displays the menu to encode, decode, and copy the text file’s data, according to the user’s needs. After the user selection is obtained, the following can be done:

- Encode
- Decode
- Copy the data

In the encoding phase, the program rearranges the plaintext using the key defined by the user. In the decoding phase, the program converts the cipher text into plaintext using the key defined by the user. In copying the data, the program copies the exact plaintext or cipher text from an input file to the target file or output file.

The program has custom methods for each task. For the encoding phase, the class has two methods:

- **encode** – this method reads the input file line by line, parses each line into words, encrypts each word character by character, and writes to the output file.
- **encodeHelper** – this method performs a subtask for the encode method: encrypt a given character according to the key specified by the user.

For the decoding operation, the class has two methods similar to those in the encoding phase:

- **decode** – similar to the *encode* method, it reads the input file line by line, parses each line into words, decrypts each word character by character, and writes to the output file.
- **decodeHelper** – this method performs a subtask for the decode method: decrypt a given character according to the key specified by the user.

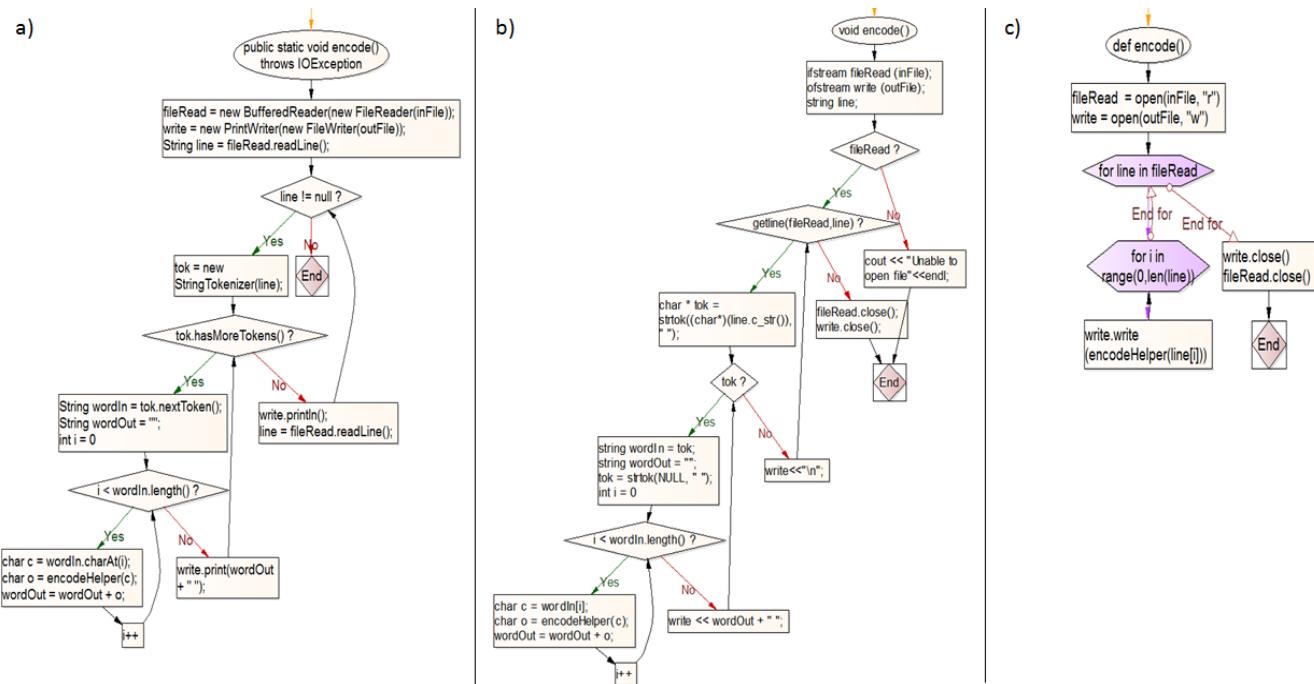


Figure 4: Code flow chart for the encode method: (a) Java, (b) C++, (c) Python

This program also has an additional method for copying one file to another, called *copy*. It copies one file to another line by line. In the following subsections, we present detailed technical descriptions for each method used in the encoding and decoding processes.

A. encode Method

This method allows for the encoding of the plaintext of the input file into the cipher text with respect to the key and with the aid of an encode helper. After the arrangement of the file read and write, the program starts reading the file text using a buffer reader line by line and storing the line in a string-type variable. We then split a string to read it word by word. After the line is split into words, the program starts to read the word character by character. After this, we get the output character with respect to the key and input character with the help of the encodeHelper method. It is important to note that the encode method performs the following steps:

1. Read a line.
2. If the line is null (i.e., end of the file), stop.
3. Break the lines into tokens by breaking around the space.
4. For each token of the line, perform the following:
 - a) Create an empty string to store the encoded word.
 - b) Get a character from the word.
 - c) Encode using the encodeHelper method.
 - d) Append to the encoded word created in step 4(a).
 - e) Repeat steps 4(b) to 4(d) for each character in the word.

5. Write the encoded word to file.

6. Repeat steps 1 to 5 for each line of the file.

To illustrate this method, assume that we have a file that contains these two lines: “Hello there. Welcome to Caesar cipher.” The method reads the file line by line; thus, line 1 is read first (“Hello there”). Then the method breaks this into tokens around a space. The two tokens are “Hello” and “there.” Then the method encodes these tokens one by one. “Hello” is encoded first. The loop encrypts the word one by one. So “Hello” is encrypted to, let us say, “Fcjjm.” This word is then written to file. Similarly, “There” is encoded and written to file. Then the program moves to read the next line and encode and write to file in the same way as in the previous line. Figure 4 illustrates the code flow chart for this method¹.

B. encodeHelper Method

The method is basically used by the encode method. It returns the encoded character after a shift to the right. The shift is done by “key” characters specified at the start of the program. The only characters supported are *a-z* and *A-Z*. The rest of the characters are returned as they are and not encoded. The following steps describe how the method works:

1. Check whether the character is *a-z* or *A-Z*.
2. If not, return the character as it is.
3. Find the new character position after shifting to the right.
4. If the character after the shifting goes out of range (becomes greater than *z*), then wrap around to start.
5. If the character is within range, return that character after the shifting.

¹All Code flow charts are available on the following link:
<https://sites.google.com/site/caesarciphermethoddesign/>

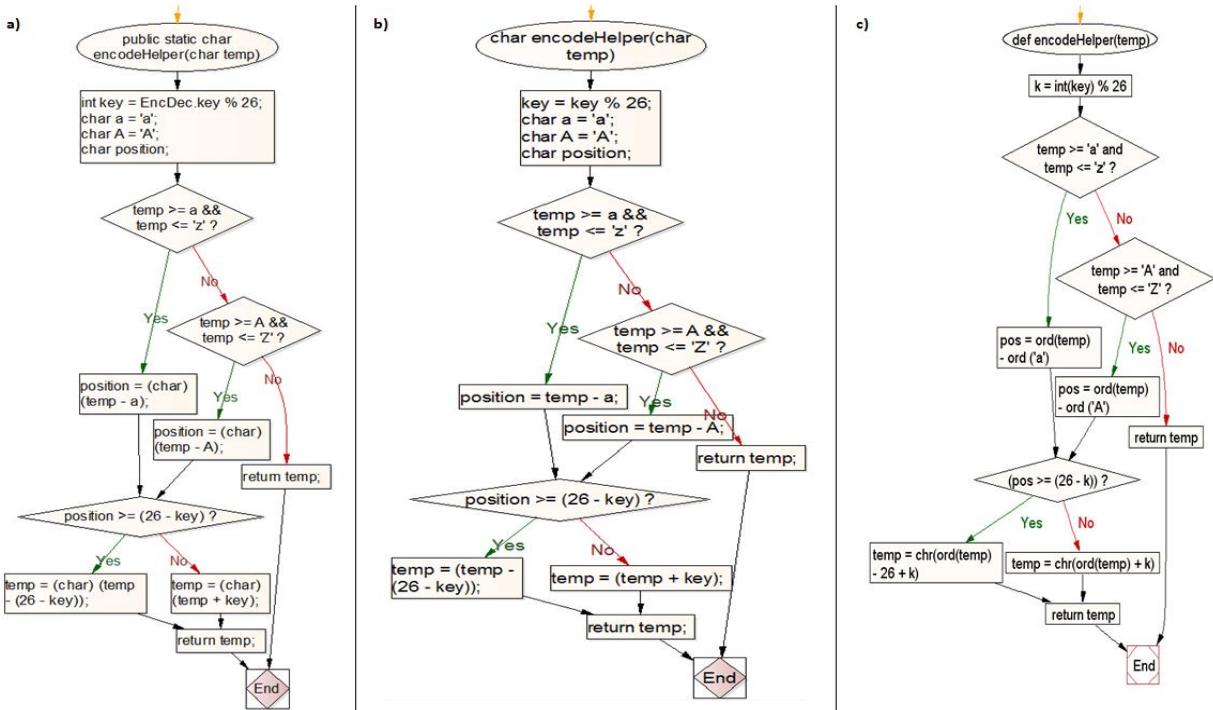


Figure 5: Code flow chart for the `encodeHelper` method: (a) Java, (b) C++, (c) Python

For example, let us assume that the value of `key = 2` and the character to encode = `c`. First, we check whether the character is `a-z` or `A-Z`, which is true. Then we find the position of the character from the start of its character set. So the position of `c` = 2 because it is two characters away from `a`. Then we check whether the position + key is greater than or equal to 26 (size of the character set). In our case, $2 + 2 < 26$; thus, we shift `c` two characters ahead, so the character that is encoded is `e`.

Now let us assume that the value of `key = 2` and the character to encode = `z`. Therefore, `z` is a supported character that can be encoded. Now the position of `z` = 25. Then we check the position + key = $25 + 2 = 27 > 26$. So we wrap around by subtracting 26 from the sum = 1. Thus, we shift by moving one character ahead of `a`. So `z` is encoded as `b`. Figure 5 describes the code flow chart for the `encodeHelper` method.

C. decode Method

This method first creates a buffered reader to read from a file and a print to write to the output file. Then the method reads from the file line by line. It is necessary to highlight that the `decode` method performs the following steps:

1. Read a line.
2. If the line is null (i.e., end of the file), stop.
3. Break the lines into tokens by breaking around the space.
4. For each token of the line, perform the following:
 - a) Create an empty string to store the decoded word.
 - b) Get a character from the word.
 - c) Encode using the `decodeHelper` method.

d) Append to the decoded word created in step 4(a).

e) Repeat steps 4(b) to 4(d) for each character in the word.

5. Write the decoded word to file.

6. Repeat steps 1 to 5 for each line of the file.

To describe this method, consider that we have a file containing the following two lines of the decrypted text:

*Fcjjmrfcp
UcjamkcrmAycqcpAgnfcpc*

The method reads this line by line; thus, line 1 is read first (*"Fcjjmrfcp"*). Then the method breaks this into tokens around a space. So the two tokens are *"Fcjjm"* and *"rfcp."* Then the method encodes these tokens one by one. So *"Fcjjm"* is decoded first. The loop decodes the word one by one. So *"Fcjjm"* is decoded to, let us say, *"Hello."* Then this word is written to file. Similarly, *"rfcp"* is decoded and written to file. Then the program moves to read the next line and decode and write to file in the same way as in the previous line. It is worthwhile to note that a description of the code flow chart for the decode method is presented in Figure 6.

D. decodeHelper Method

This method returns the decoded character after a shift to the left. The shift is done by the "key" characters specified at the start of the program. The only characters supported are `a-z` and `A-Z`. The rest of the characters are returned as they are and not decoded. The following important steps show how the method works:

1. Check whether the character is `a-z` or `A-Z`.
2. If not, return the character as it is.

3. Find the new character position after shifting to the left.
4. If the character goes out of range (becomes less than a or A), wrap around to the end of the character set.
5. If the character is within range, return that character after shifting to the left.

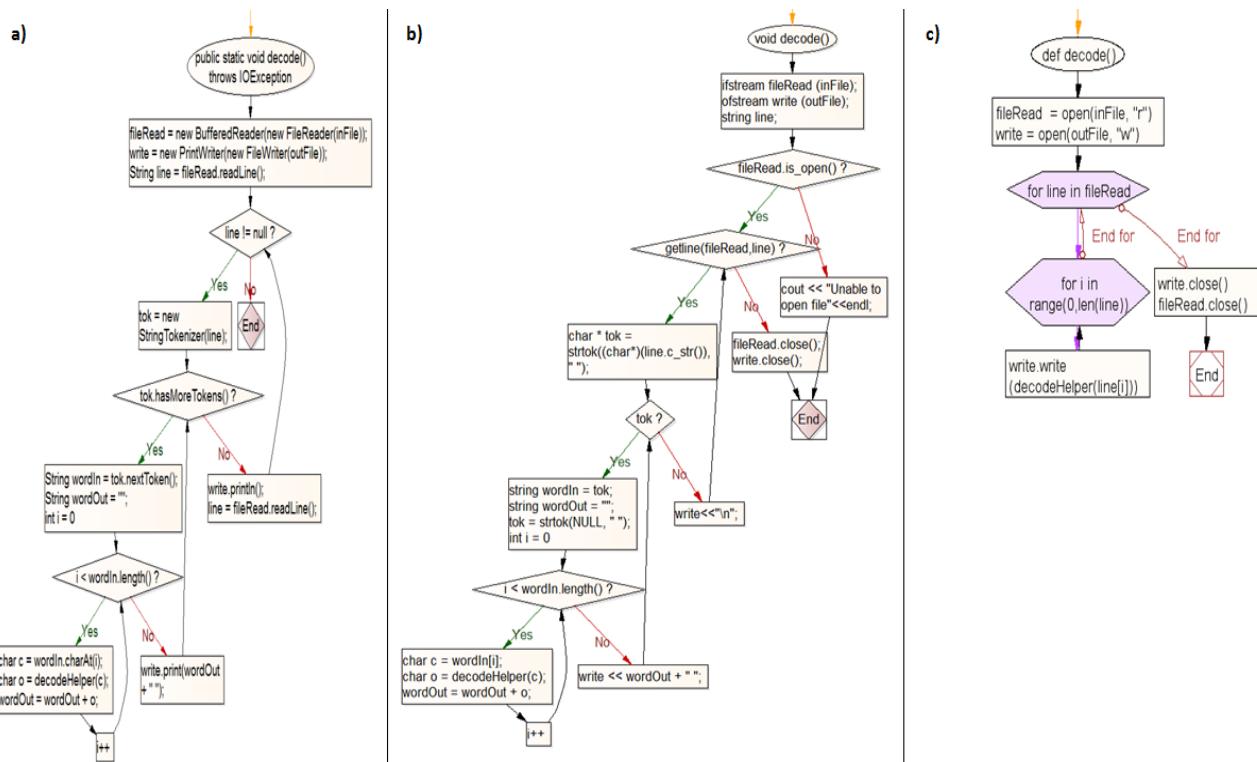


Figure 6: Code flow chart for the decode method: (a) Java, (b) C++, (c) Python

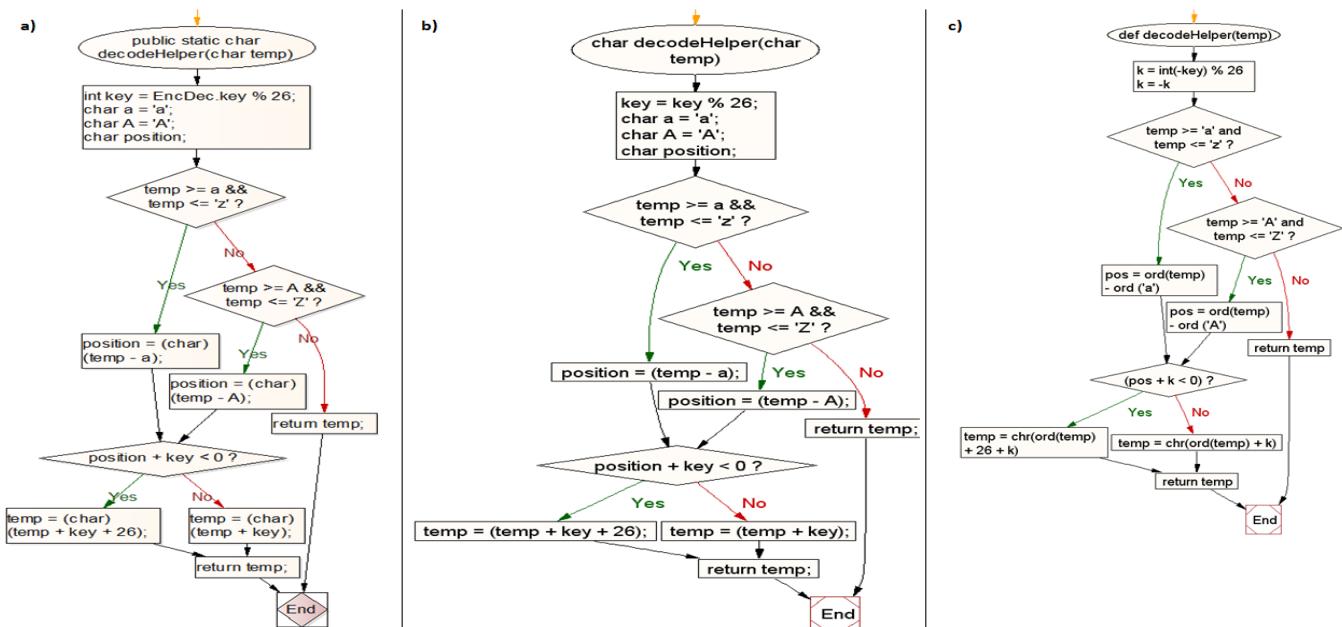


Figure 7: Code flow chart for the decodeHelper method: (a) Java, (b) C++, (c) Python

Now let us assume that the value of key = -2 and the character to decode = *a*. Thus, *a* is a supported character that can be decoded. Now the position of *a* = 0. Then we check the position + key = 0 - 2 = -2 < 0. So we wrap around by adding 26 to the sum = -2 + 26 = 24. Thus, we shift by moving 24 characters ahead of *a*. Therefore, *a* is decoded as *y*.

IV. CONCLUSION

Cryptography deals with various techniques that can be used to convey information in a secure fashion. Its objective is to enable the recipient of the message to receive it properly and stop eavesdroppers from understanding it. Cryptography is the art and science of turning an original message into a form that is completely unreadable. The two techniques used to convert data into an unreadable form are the transposition technique and the substitution technique. The Caesar cipher uses the substitution method.

The Caesar cipher algorithm is one of the oldest algorithms. Far newer algorithms that are far more secure have emerged, but the Caesar cipher algorithm is still the fastest because of its simplicity. But it is very easy to crack.

We implemented three programs in this paper based on Java, C++, and Python to implement the Caesar cipher encryption algorithm to aid information security students and help them understand this algorithm. Therefore, it is possible to use this paper for educational purposes in the field of information and communication security. It is crucial to highlight that a code flow chart is used for each program, and a chart is also used to describe the flow of the code. This further reveals the step sequence for major methods used in the code and the relationships between them. Some detailed technical descriptions are also presented for each method used to encode and decode messages. Furthermore, the weaknesses and limitations of the classical Caesar cipher are clearly described.

APPENDIX A. JAVA CODE

```

import java.io.*;
import java.util.*;

public class EncDec {
    static BufferedReader stdin = new BufferedReader(new InputStreamReader(System.in));
    static PrintWriter write;
    static BufferedReader fileRead;
    static String inFile, outFile;
    static int key;
    static StringTokenizer tok;

    public static void main(String[] args) throws IOException {
        System.out.print("Welcome to File Encryption/Decryption Program");
        System.out.println("*****");
        System.out.println();
        System.out.print("Enter Source File Name> ");
        inFile = stdin.readLine();
        System.out.print("Enter Target File Name> ");
        outFile = stdin.readLine();
        System.out.print("Enter Key (>0: encryption, <0 decryption, 0: copying)> ");
        key = Integer.parseInt(stdin.readLine());

        if (key > 0) {
            encode();
        } else if (key < 0) {
            decode();
        } else {
            copy();
        }
        System.out.println(inFile + " " + outFile + " " + key);
        System.out.println(".....");
        System.out.println("Done, please check output file");
        write.close();
    }

    public static void encode() throws IOException {
        fileRead = new BufferedReader(new FileReader(inFile));
        write = new PrintWriter(new FileWriter(outFile));
        String line = fileRead.readLine();
        while (line != null) {
            tok = new StringTokenizer(line);
            while (tok.hasMoreTokens()) {
                String wordIn = tok.nextToken();
                String wordOut = "";
                for (int i = 0; i < wordIn.length(); i++) {
                    char c = wordIn.charAt(i);
                    char o = encodeHelper(c);
                    wordOut = wordOut + o;
                }
                write.print(wordOut + " ");
            }
            write.println();
            line = fileRead.readLine();
        }
    }

    public static void decode() throws IOException {
        fileRead = new BufferedReader(new FileReader(inFile));
        write = new PrintWriter(new FileWriter(outFile));
        String line = fileRead.readLine();
        while (line != null) {
            tok = new StringTokenizer(line);
            while (tok.hasMoreTokens()) {
                String wordIn = tok.nextToken();
                String wordOut = "";
                for (int i = 0; i < wordIn.length(); i++) {
                    char c = wordIn.charAt(i);
                    char o = decodeHelper(c);
                    wordOut = wordOut + o;
                }
                write.print(wordOut + " ");
            }
            write.println();
            line = fileRead.readLine();
        }
    }
}

```

```

public static void copy() throws IOException {
    fileRead = new BufferedReader(new FileReader(inFile));
    write = new PrintWriter(new FileWriter(outFile));
    String line = fileRead.readLine();
    while (line != null) {
        write.println(line);
        line = fileRead.readLine();
    }
}

public static char encodeHelper(char temp) {
    int key = EncDec.key % 26; //added this line to support any key
    char a = 'a';
    char A = 'A';
    char position;
    if (temp >= a && temp <= 'z')
        position = (char) (temp - a);
    else if (temp >= A && temp <= 'Z')
        position = (char) (temp - A);
    else
        return temp;
    if (position >= (26 - key))
        temp = (char) (temp - (26 - key));
    else
        temp = (char) (temp + key);
    return temp;
}

public static char decodeHelper(char temp) {
    int key = EncDec.key % 26; //added this line to support any key
    char a = 'a';
    char A = 'A';

    //Changed how to calculate position
    char position;
    if (temp >= a && temp <= 'z')
        position = (char) (temp - a);
    else if (temp >= A && temp <= 'Z')
        position = (char) (temp - A);
    else
        return temp;

    //changed how to check for out of range
    if (position + key < 0)
        temp = (char) (temp + key + 26);
    else
        temp = (char) (temp + key);
    return temp;
}
}

```

APPENDIX B. C++ CODE

```

// Libraries
#include <string>
#include <vector>
#include <fstream>
#include <iostream>
using namespace std;
static string inFile;
static string outFile;
int key = 0;
// Encode Helper function used in encode function
char encodeHelper(char temp)
{
    // Added this line to support any key
    key = key % 26;
    char a = 'a';
    char A = 'A';
    char position;
    if (temp >= a && temp <= 'z')
    {
        position = temp - a;
    }
    else if (temp >= A && temp <= 'Z')
    {
        position = temp - A;
    }
    else
    {
        return temp;
    }

    if (position >= (26 - key))
    {
        temp = (temp - (26 - key));
    }
    else
    {
        temp = (temp + key);
    }
    return temp;
}
// Decode Helper function used in Decode function
char decodeHelper(char temp)
{
    // Added this line to support any key
    key = key % 26;
    char a = 'a';
    char A = 'A';

    // Changed how to calculate position
    char position;
    if (temp >= a && temp <= 'z')

```

```

{
    position = (temp - a);
}
else if (temp >= A && temp <= 'Z')
{
    position = (temp - A);
}
else
{
    return temp;
}

// Changed how to check for out of range
if (position + key < 0)
{
    temp = (temp + key + 26);
}
else
{
    temp = (temp + key);
}
return temp;
}

// Encode function used for Encoding
void encode()
{
    ifstream fileRead (inFile);
    ofstream write (outFile);
    string line;
    // If file exists
    if (fileRead)
    {
        // Read file line by line
        while (getline(fileRead,line))
        {
            // Tokenize line
            char * tok = strtok((char*)(line.c_str()), " ");
            while (tok)
            {
                {
                    string wordIn = tok;
                    string wordOut = "";
                    tok = strtok(NULL, " ");
                    // Encode each character of each token using
                    encodeHelper function
                    for (int i = 0; i < wordIn.length(); i++)
                    {
                        char c = wordIn[i];
                        char o = encodeHelper(c);
                        wordOut = wordOut + o;
                    }
                    write << wordOut + " ";
                }
                write<<"\n";
            }
            fileRead.close();
            write.close();
        }
    }
    else
    {
        cout << "Unable to open file" << endl;
    }
}

// Decode function used to Decode files
void decode()
{
    ifstream fileRead (inFile);
    ofstream write (outFile);
    string line;
    if (fileRead.is_open())
    {
        // Read file to be decoded line by line
        while (getline(fileRead,line))
        {
            // Tokenize line
            char * tok = strtok((char*)(line.c_str()), " ");
            while (tok)
            {
                {
                    string wordIn = tok;
                    string wordOut = "";
                    tok = strtok(NULL, " ");
                    // Decode each character of each token
                    for (int i = 0; i < wordIn.length(); i++)
                    {
                        char c = wordIn[i];
                        char o = decodeHelper(c);
                        wordOut = wordOut + o;
                    }
                    write << wordOut + " ";
                }
                write<<"\n";
            }
            fileRead.close();
            write.close();
        }
    }
    else
    {
        cout << "Unable to open file" << endl;
    }
}

// Copy function used for Copying
void copy()
{
    ifstream fileRead (inFile);
    ofstream write (outFile);
    string line;
    if (fileRead.is_open())
    {
        while (getline(fileRead,line))
        {
            write << line << endl;
        }
        fileRead.close();
        write.close();
    }
    else
    {
}

```

```

        cout << "Unable to open file" << endl;
    }
}

// Main Program
void main()
{
    cout << "Welecome to File Encryption/Decryption Program" << endl;
    cout << "*****" << endl;
    cout << endl;
    cout << "Enter Source File Name> ";
    getline (cin,inFile);
    cout << "Enter Target File Name> ";
    getline (cin,outFile);
    cout << "Enter Key (>0: encryption, < 0 decryption, 0: copying)> ";
    cin>>key;

    if (key > 0)
    {
        encode();
    }
    else if (key < 0)
    {
        decode();
    }
    else
    {
        copy();
    }
    cout << inFile << " " << outFile << " " << key << endl;
    cout << "....." << endl;
    cout << "Done, please check output file" << endl;
}

```

APPENDIX C. PYTHON CODE

```

global inFile
global outFile
global key

def encode():
    fileRead = open(inFile, "r")
    write = open(outFile, "w")
    for line in fileRead:
        for i in range(0,len(line)):
            write.write(encodeHelper(line[i]))
    write.close()
    fileRead.close()

def encodeHelper(temp):
    k = int(key) % 26

    if temp >= 'a' and temp <= 'z':
        pos = ord(temp) - ord ('a')
    elif temp >= 'A' and temp <= 'Z':
        pos = ord(temp) - ord ('A')
    else:
        return temp

    if(pos >= (26 - k)):
        temp = chr(ord(temp) - 26 + k)
    else:
        temp = chr(ord(temp) + k)

    return temp

def decode():
    fileRead = open(inFile, "r")
    write = open(outFile, "w")
    for line in fileRead:
        for i in range(0,len(line)):
            write.write(decodeHelper(line[i]))
    write.close()
    fileRead.close()

def decodeHelper(temp):
    k = int(-key) % 26
    k = -k

    if temp >= 'a' and temp <= 'z':
        pos = ord(temp) - ord ('a')
    elif temp >= 'A' and temp <= 'Z':
        pos = ord(temp) - ord ('A')
    else:
        return temp

    if(pos + k < 0):
        temp = chr(ord(temp) + 26 + k)
    else:
        temp = chr(ord(temp) + k)

    return temp

def copy():
    fileRead = open(inFile, "r")
    write = open(outFile, "w")
    write.write(fileRead.read())
    write.close()
    fileRead.close()

print "Welcom to File Encryption/Decryption Program"
print "*****\n"
print "Enter Source File Name> "
inFile = 'x' # raw_input()
print "Enter Target File Name> "
outFile = 'a' # raw_input()
print "Enter Key (>0: encryption, < 0 decryption, 0: copying)> "
key = 0 # raw_input()

if key > 0:
    encode()

```

```

    elif key < 0:
        decode()
    else:
        copy()

    print inFile + " " + outFile + " " + str(key)
    print "....."
    print "Done, please check output file"

```

REFERENCES

- [1] Eskicioglu, A., & Litwin, L. (2001). Cryptography. *IEEE Potentials*, 20(1), 36-38.
- [2] Udo, G. J. (2001). Privacy and security concerns as major barriers for e-commerce: a survey study. *Information Management & Computer Security*, 9(4), 165-174.
- [3] Chellappa, R. K., & Pavlou, P. A. (2002). Perceived information security, financial liability and consumer trust in electronic commerce transactions. *Logistics Information Management*, 15(5/6), 358-368.
- [4] Goldreich, O. (2005). Foundations of Cryptography—A Primer. *Foundations and Trends® in Theoretical Computer Science*, 1(1), 1-116.
- [5] Kahate, A. (2013). *Cryptography and network security*. Tata McGraw-Hill Education.
- [6] Calabrese, T. (2004). *Information security intelligence: Cryptographic principles and applications*. Cengage Learning.
- [7] Lubbe, J. C. (1998). *Basic methods of cryptography*. Cambridge University Press.
- [8] Stallings, W. (2006). *Cryptography and network security: principles and practices*. Pearson Education India.
- [9] Ahmad, M., Khan, I. R., & Alam, S. (2015). Cryptanalysis of image encryption algorithm based on fractional-order lorenz-like chaotic system. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2* (pp. 381-388). Springer International Publishing.
- [10] Bishop, D. (2003). Introduction to cryptography with Java applets. Jones & Bartlett Learning.
- [11] Menezes, A. J., Van Oorschot, P. C., & Vanstone, S. A. (1996). *Handbook of applied cryptography*. CRC press.
- [12] Dooley, J. F. (2013). *A brief history of cryptology and cryptographic algorithms*. Springer.
- [13] Manasrah, A. M., & Al-Din, B. N. (2016). Mapping private keys into one public key using binary matrices and masonic cipher: Caesar cipher as a case study. *Security and Communication Networks*.
- [14] Mishra, A. (2013). Enhancing security of caesar cipher using different methods. *International Journal of Research in Engineering and Technology*, 2(09), 327-332.
- [15] Gowda, S. N. (2016). Innovative enhancement of the Caesar cipher algorithm for cryptography. In *Advances in Computing, Communication, & Automation (ICACCA)(Fall)*, International Conference on (pp. 1-4). IEEE.
- [16] Robling Denning, D. E. (1982). *Cryptography and data security*. Addison-Wesley Longman Publishing Co., Inc.
- [17] Knudsen, L. R. (1998). Block Ciphers—a survey. In *State of the Art in Applied Cryptography* (pp. 18-48). Springer Berlin Heidelberg.
- [18] Salomon, D. (2003). Introduction. In *Data Privacy and Security* (pp. 1-17). Springer New York.
- [19] Omolara, O. E., Oludare, A. I., & Abdulahi, S. E. (2014). Developing a modified Hybrid Caesar cipher and Vigenere cipher for secure Data Communication. *Computer Engineering and Intelligent Systems*, 5, 34-46.
- [20] Govinda, K. (2011). Multilevel cryptography technique using graceful codes. *Journal of Global Research in Computer Science*, 2(7), 1-5.
- [21] Bruen, A. A., & Forcinito, M. A. (2011). *Cryptography, information theory, and error-correction: a handbook for the 21st century* (Vol. 68). John Wiley & Sons.
- [22] Jain, A., Dedhia, R., & Patil, A. (2015). Enhancing the security of caesar cipher substitution method using a randomized approach for more secure communication. *arXiv preprint arXiv:1512.05483*.
- [23] Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016). Privacy-preserving-outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security*, 11(8), 1847-1861.

AUTHOR PROFILE

Ismail Keshta is an assistant professor at the Department of Computer and Information Technology at Dammam Community College (DCC), King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. He received the B.Sc. and M.Sc. degrees in computer engineering and the Ph.D. degree in computer science and engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2009, 2011, and 2016, respectively. He was a Lecturer with the Computer Engineering Department, KFUPM, from 2012 to 2016. Prior to that, in 2011, he was a Lecturer with Princess Nora Bint Abdulrahman University (PNU) and Imam Muhammad ibn Saud Islamic University (IMAMU), Riyadh, Saudi Arabia. His research interests include software process improvement, modeling, and intelligent systems.

Chronic Kidney Disease Prediction Using Machine Learning

Sathiya Priya S
PG Scholar,

Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Suresh Kumar M
Professor,

Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Abstract - Chronic Kidney Disease prediction is one of the most important issues in healthcare analytics. The most interesting and challenging tasks in day to day life is prediction in medical field. In this paper, we employ some machine learning techniques for predicting the chronic kidney disease using clinical data. We use three machine learning algorithms such as Decision Tree(DT) algorithm, Naive Bayesian (NB) algorithm. The performance of the above models are compared with each other in order to select the best classifier in predicting the chronic kidney disease for given dataset.

Index Terms – Machine Learning; Chronic Kidney Disease; Prediction.

1. INTRODUCTION

Computer vision has been one of the most remarkable breakthroughs for the machine learning and in particular for active healthcare applications. Machine learning allows to build the models to quickly analyze data and deliver results for the given data. Healthcare service providers can make better decisions on patient's disease diagnosis and treatment for the particular disease with the help of machine learning. The massive quantities of data are analysed using machine learning. It delivers faster and more accurate results in order to identify the risks, it may also require additional time and resources to train it proper manner.

Supervised machine learning algorithms can applied to predict the future events with the help of what has been learned in the past to new data using labeled examples. First the known training dataset is analyzed, with that the learning algorithm produces an inferred function to make predictions about the output values.

After sufficient training the system is able to provide targets for any new inputs. Supervised learning algorithms uses patterns to predict label values on additional unlabeled data. As per Fig 1.1 Machine learning algorithms are classified in two types they are supervised machine learning algorithms and unsupervised machine learning algorithms. Supervised machine learning algorithms are based on input-output pairs patterns [1]. These algorithms aims to predict output values based on given input values. Supervised machine learning algorithms mainly focuses on classification and regression.

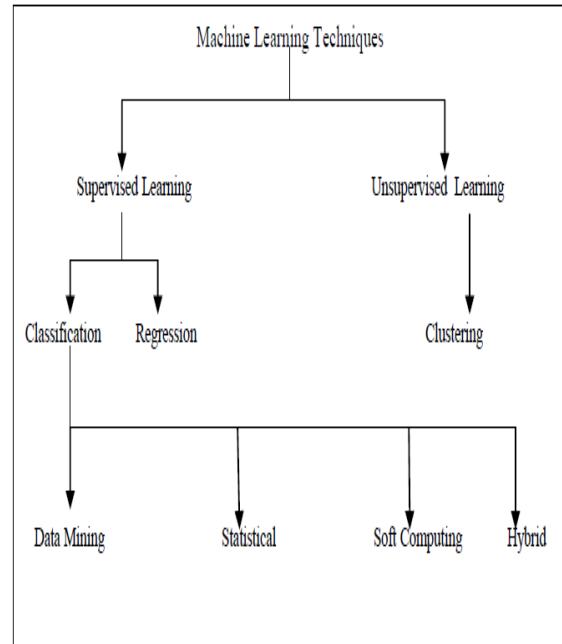


Fig 1.1 classification of machine learning techniques.

2. RELATED WORK

Chronic kidney disease will harm kidneys and reduce its ability to keep our body in a healthy condition. The risk of having heart and blood vessel disease increases due to kidney disease. For the considered dataset the Random Forest has produced better prediction performance in terms of classification accuracy, AUC respectively. The classification performances of the classifier is analyzed with the standard performance parameters, such as: Accuracy, Specificity, Sensitivity, Precision [2].

The machine learning algorithms behaviour were determined on a set of data mining indicators has a relative effect on the models. Knowledge discovery from the wide databases is known as Data mining. Besides studying the existing available Clinic Foundation Heart Disease dataset, 600 clinical records collected from a leading Chennai based diabetes research centre. The application of Data mining technique is a good method for different analysis of medical data. [3]. The chronic disease is predicted with the clinical data using machine learning algorithms. The machine learning algorithms used here includes K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR). The predictive models are constructed for the taken dataset and the best classifier is predicted using the performance of the models. SVM classifier gives the highest accuracy and has the highest sensitivity after training and testing [4]. The healthcare industry collects large amounts of medical data and for effective decision making the data need to be mined to discover hidden information. Based on the clinical data of patients the heart disease prediction system can assist medical professionals in the prediction of heart disease. The marginal success is achieved with the predictive model for heart disease patients and there is a need for more complex models to increase the accuracy of prediction of the early stages of heart disease [5]. Classification technique normally divides the data into two different data sets one is training set and the other is testing sets. Every occurrence in the training set contains one target variable and several attributes or features. The training data is used to develop a model in SVM, their features which successively predicts the target values of the test data given only the attributes of the input test data. Random Forest is a collection of a group of tree predictors which uses classification technique [16].

3. PROPOSED SYSTEM

The proposed system deals with the prediction of chronic disease from the clinical data. The healthcare generates large data, so it is necessary to collect this data and effectively use it for analysis, prediction, and treatment.

A classification model draws some conclusion from observed values. In classification model one or more inputs are used to predict the value of one or more outcomes. The dataset is applied with the labels which are outcomes. In a supervised machine learning algorithms, the classification algorithm uses the training dataset. classification predicts the categorical class labels whereas the prediction predicts the unknown or missing values.

A decision tree is a tree structure in which internal nodes i.e., non leaf nodes denotes a test on an attribute. Branches denotes the outcomes of tests. Leaf nodes i.e., terminal nodes hold class labels. Root node is the topmost node in the decision tree. A path is traced to leaf node from root node which holds the prediction for the given tuple.

Any domain knowledge or parameter setting is not required for the construction of a decision tree. Decision tree can handle high dimensional data and it can be understood by humans easily. Learning and classification are simple ,fast and it has good accuracy.

Naive Bayes classifier is a powerful algorithm for the classification task. Even with working on a data set with millions of records with some attributes, Naive Bayes approach is best to use. Naive Bayes classifier uses the Bayes Theorem. For each class it predicts membership probabilities such as the probability that given record or data point belongs to a particular class. ‘A’ denotes prior event and ‘B’ denotes dependent event, Bayes’ theorem can be given as

$$\text{Prob(A|given B)} = \text{Prob}(A \text{and } B) / \text{Prob}(B)$$

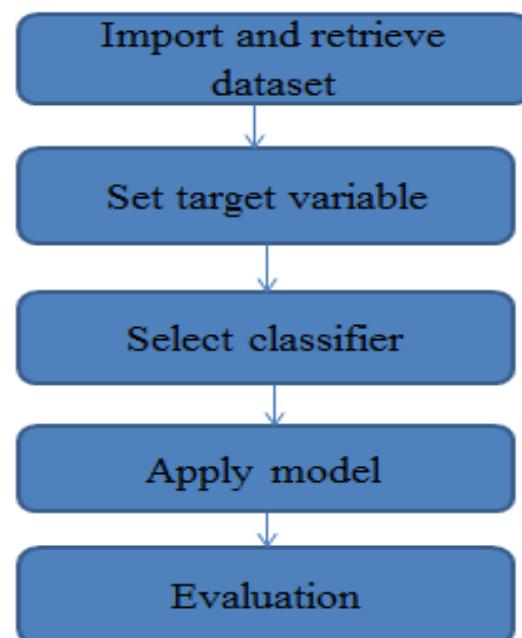


Fig 3.1 The process model to predict chronic kidney disease.

The process steps for Fig 3.1 is as follows: First the dataset is imported and retrieved using the basic steps. Then, the target variable is fixed. Next, the classification method is selected to predict the chronic kidney disease.

Then, the model for the classifier is applied to get the prediction results. Finally, the results of the different classifier are evaluated using some parameters.

4. DATASET

The proposed system uses the dataset taken from the UCI Machine Learning Repository named Chronic Kidney Disease has 25 attributes, 11 numeric and 14 nominal. Total 400 instances of the dataset is used for the training to prediction algorithms, out of which 250 has label chronic kidney disease (CKD) and 150 has label non chronic kidney disease (NOTCKD). The attributes in the dataset are age, bp, sg, al, su, bc, pc, pcc, ba, bgr, bu, sc, sod, pot, hemo, pcr, wc, rc, htn, dm, cad, appet, pe, ane, classification. The dataset is divided into two groups, one for training and another for testing. The ratio of training and testing data is 70% and 30% respectively.

5. RESULT AND DISCUSSION

The machine learning methods described are trained to predict the chronic kidney disease. Two classifier methods are used in this decision tree and naive bayes . The experiments are constructed on R tool. In this work , the performance is measured by sensitivity, specificity and accuracy described as follows.

Accuracy (ACC) is the overall success rate of the classifier defined as

$$ACC = (TP + TN) / (TP + FP + TN + FN)$$

Sensitivity or the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model defined as

$$Sensitivity = TP / (TP + FN).$$

Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model defined as

$$Specificity = TN / (FP + TN).$$

Where

TP - the number of true positives.

TN - the number of true negatives.

FP - the number of false positives.

FN - the number of false negatives.

With the help of True Positive (TP) and True Negative (TN) the performance of the classifications model is evaluated. The machine learning techniques used are trained and tested separately in this work. The 10-fold cross validation is used to train and test the machine learning models in this work and the average results are shown in table 5.1.

Table 5.1 Performance evaluation for Decision Tree and Naive Bayes classification techniques.

Techniques used	Accuracy	Sensitivity	Specificity
Decision Tree	99.25%	99.20%	99.33%
Naive Bayes	98.75%	98%	98.75%

From table 5.1 by comparing the decision tree method and naive bayes method the accuracy of decision tree method is relatively higher than the naive bayes method. The decision tree method can be adopted since it has the accuracy of 99.25% in prediction of chronic kidney disease.

6. CONCLUSION

The prediction of chronic kidney disease is very important and now-a-days it is the leading cause of death. The performance of Decision tree method was found to be 99.25% accurate compared to naive Bayes method. Classification algorithm on chronic kidney disease dataset the performance was obtained as 99.33% Specificity and 99.20% Sensitivity. We are also further working on enhancing the performance of prediction system accuracy in neural network and deep learning algorithm .

7. REFERENCES

- [1] Madhura Rambajani, Wyomesh Deepanker, Neelam Pathak (2015), “A Survey On Implementation Of Machine Learning Techniques For Dermatology Diseases Classification”, International Journal of Advances in Engineering & Technology.
- [2] Manish Kumar (2016), “Prediction Of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm”, International Journal of Computer Science and Mobile Computing , Vol. 5, Issue. 2, pg.24 – 33.

- [3] K. R. Anantha Padmanaban and G. Parthiban (2016), “Applying Machine Learning Techniques For Predicting The Risk Of Chronic Kidney Disease” Indian Journal of Science and Technology, Vol. 9(29).
- [4] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, (2016) “predictive analytics for chronic kidney disease using machine learning techniques”, The 2016 Management and Innovation Technology International Conference .
- [5] Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel, (2016) “Heart Disease Prediction Using Machine learning and Data Mining Technique”, International Journal Of Computer Science & Communication, Vol. 7, No. 1, pp.129 – 137.
- [6] Jerez, J. M., Molina, I., Garcia-Laencina, P. J., Alba, E., Ribelles, N., Martin, M., and Franco, L, (2010) “Missing data imputation using statistical and machine learning methods in a real breast cancer problem”. Artif. Intell., Med. 50, 2, 11–11.
- [7] A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository [Online]. Available: <http://www.ics.uci.edu/~mlearn/MLRepository.html>
- [8] Witten H, Ian H, (2011) “Data mining: practical machine learning tools and techniques”, Morgan Kaufmann Series in Data Management Systems.
- [9] P. B. Jensen, L. J. Jensen, and S. Brunak, (2012) “Mining electronic health records: towards better research applications and clinical care,” Nature Reviews Genetics, Vol. 13, no. 6, pp. 395–405.
- [10] J. C. Ho, C. H. Lee, and J. Ghosh , (2014) “Septic shock prediction for patients with missing data,” ACM Transactions on Management Information Systems (TMIS), Vol. 5, no. 1, p. 1.
- [11] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, (2015) “A relative similarity based method for interactive patient risk prediction,” Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093.
- [12] Hudson Fernandes Golino, Liliany Souza de Brito Amaral, Stenio Fernando Pimentel Duarte, Cristiano Mauro Assis Gomes, Telma de Jesus Soares, Luciana Araujo dos Reis, and Joselito Santos, (2014) “Predicting increased blood pressure using machine learning”, Journal of obesity.
- [13] Tina Patil, R & Sherekar, SS, (2013) “Performance Analysis of Naive bayes and J48 Classification Algorithm for Data Classification”, International Journal of Computer Science and Applications, vol. 6, no.2, pp. 256-261.
- [14] D. M. F. bin Othman and T. M. S. Yau, (2007) “Comparison of Different Classification Techniques Using WEKA for Breast Cancer,” in 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, F. Ibrahim, N. A. A. Osman, J. Usman, and N. A. Kadri, Eds. Springer Berlin Heidelberg, pp. 520–523.
- [15] Taiwo Oladipupo Ayodele, (2010) “Types of Machine Learning Algorithms”, New Advances in Machine Learning, Yagang Zhang (Ed.), InTech.
- [16] Ashfaq Ahmed K, Sultan Aljahdali, Nisar Hundewale and Ishthaq Ahmed K , “Cancer Disease Prediction With Support Vector Machine And Random Forest Classification Techniques”, IEE Cybernetics 2012.

Learning engagement based on cloud classroom

Nasrin Akter

Computer Science and Engineering
Daffodil International University
nasrin.cse@diu.edu.bd

Shah Akbar Ahmad

Computer Science and Engineering
Daffodil International University
shuvo2220@diu.edu.bd

ABSTRACT

Present day showing techniques request imaginative and powerful utilization of innovation at most extreme level. Consolidating a virtual group outside classroom instructing has turned out to be inescapable in computerized age training. This exploration was planned to discover how this can be utilized as a part of terms of intuitive instructing and how it can encourage understudies to recuperate the absences of learning in classroom. A web group of a college called Learning Feedback System (LFS) has been utilized here as the strategy to break down five example cases. Impacts of A critical level of connection in LFS showed that it decreased the correspondence hole between understudies and educators that obviously prompting appropriate learning.

Keywords

Interactive teaching, Learning feedback method, virtual community, technology education.

1. INTRODUCTION

Technology transfusion have evermore had an exigent dominance on industry upliftment, touching even the most consecutive systems such as study. Following the ordinary mutation of people's exercise and the world's job market composition, the education segment has gone per a massive association metamorphosis over the dernier several years. The tools allow for shared entrance to teaching components and support education-base information. When it appears to digital learning, there are a numerous tools and solicitation to pick from. Which ones choose

to use should really be about what works best for educational needs, over the loyalty to a particular brand.

The thought of virtual classroom has its variant illustration solicitation, but it for starters allude to technology that inflict potential possessing via the web. The amenities of proposing these scheme are most repeatedly observed in kinship to business, but its impression on the learning segment is no under momentous. Educational establishment in the universe have meanwhile adapted the cube to their radical appointment and formed use of its august brawny for newness.

2. LITERATURE REVIEW

Sometime before the invention of modern communication technology such as Computers, internet, Mathematician from the Massachusetts Institute of Technology, also a computer scientist, and an education visionary Seymour Paper found the idea very worth investigating that, the connected electronic device can be used to improve the experience of student while they are learning. Even the students from poverty line could be benefited from it even they are isolated geographically and socially. [1] Bangladesh is an inspiring country. Bangladeshi government declared the priority of ICT sector as usage of the eye of "Digital Bangladesh". They are supporting the corporation of technology into different equilibrium of education including the school, college and universities.

Only lectures can't fulfill the strong leaning for deferent type of students in the classroom. [2] There are many technologies (online discussion, video, audio etc) can be the ways

of instructing and instruction process. With this process many institutions are enhancing different type of method for better solution of the students. But now and then it can be listen to explain each class conversation and there are some other restriction. In that case, the virtual classroom can be a high opportunity to help the students' discussion with their teachers and friends also. Procedure for Paper Submission.

3. PROBLEM IN CLASSROOM

It is often difficult for a teacher to reach all the students and discuss with their problems in a limited hour class. [3] For the all teachers, it is often hard to gratify all the students to rehearse them. Again For all the students, it is hourly very ticklish to feel all the class lecture and ask question to their teacher in limited time. If we dispose a big size class for individual counseling then it may be not a good solution for every time and can't be practical. "There are many learners who are gawky and do not disinterest to conversation with their teacher". There are divers' ways to create the classroom usable and amusing for something like conversation, playing and many other avenue to create the classroom winsome. It is not guarantee that 100% satisfaction of the students. So the teachers can not 100% monitoring of their students.

There are abundant students who are pass there H S C exam in "Bangla" version. When they admit their university then it is more difficult to learn their superior education in English version. Most of the students can't peruse and scribe English properly. They can't write their answer in exam hall clearly. Sometimes they can't capable to understand their question paper. Yousuf Islam conducted an investigation on the matter and the result that he found is enclosed here: "To understand the trouble students have with sentence making, 18 students in first year, first semester were given a task in free-writing –any incident in their lives that affected them. The 334 collective sentences were analyzed for the type of mistakes made. 63.5% of the sentences were found to have one or more problems. The highest type of error was subject-verb

matching which was found in 19.3% of the sentences followed by preposition errors at 14.2%, etc. Students who wrote correct sentences were generally found to be writing simple sentences like those of a 3rd or 4th grader. This whole exercise was abandoned when it was accidentally discovered that when a student was asked to repeat the whole exercise, the types of mistakes were different. After a few trials, it was concluded that errors are randomly made –students have little or no idea of sentence construction [4]".

4. WEB TECHNOLOGY IN THE CLASSROOM

In 21st century, technology in the classroom is becoming more and more rising. Many other devices like tablets are replacing textbooks. Social media has become usual. We can use technology has completely transformed the way we live or lives. Students believe and prefer that technology can make learning more fun and interesting. Students deem boring or challenging can become more interesting with virtual education, through a status, file, video or when using a web technologies. Many educators argues that the only way to continue our dominance and prosperity in the world economy and politics is to educate our people as competitive and creative members of the global community, and the proper integration of technology in our education system is crucial to accomplish the goal. Numerous research papers, articles, and books were written on integrating technology into the classroom, and they often couple it with the 'constructivist' learning theory. Many of them using technology to enhance teacher-student communication and to promote collaborative and active learning, which calls for a dramatic paradigm shift from the lecture based education model. [5]

Some students are embarrassed or shy when they are called on in class. An alternative is to call on 2 people, perhaps sitting beside each other, together. Collectively they might do a better job than asking 2 students separately. This technique works especially well if the students have to think about an answer or do a calculation, or

work at the board. This tip was suggested by Marion Cohen, a part time math faculty. [6]

Technology also play a big role in the relationship between the students and the teachers: When technology is effectively integrated into subject areas, teachers grow into roles of guides and facilitator's. While students take responsibility for their learning outcomes, Technology lends itself as the multidimensional tool that assists that process. And help to makes teaching and learning process more meaningful and fun. [7]

Technology can have a reciprocal relationship with teaching. The emergence of new technologies pushes educators to understanding and leveraging these technologies for classroom use; at the same time, the on-the-ground implementation of these technologies in the classroom can (and does) directly impact how these technologies continue to take shape.[8] There are many ways to integrated technology in the classroom. Many teachers and students don't know how they use technology in their classroom. The Education World Tech Team offers lessons and activities to help educators make better use of technology tools for instruction, and to help students improve their technology skills within the context of the regular curriculum. Included: Integration activities that utilize the Web, PowerPoint, Excel, digital photography, SMART Boards, and more.[9] There are also many educators, less familiar and less comfortable with web technology then there students. So our virtual classroom help you make the best use of web technology in your school, collage and

universities. We are sharing some many user friendly features. Using this web technology in the classroom will be easier for the teachers and students.

5. METHODOLOGY

This research paper will discuss of web technology that added in the virtual classroom. And also added how teachers and students benefited can. The researcher observed other web technology tools that has already integrated in the classroom.

Course Blog and Wiki spaces allow students to hold online written discussions. Both of these promote editing. Blog allows the writer to post commentary and visitors to leave responses whereas wiki permits the live editing and revision of content of posted document [10].

This research paper will be completed in a professional manner. Here the researcher destination which encourages the collaboration between teacher and student. Here the environment is paper free. In this virtual web technology classroom is to be in the improvement of educator's workflow and therefore saving much-needed time.

The major motives of the experiment paper are mentioning bellow:

- Interface should be light to use.
- Spread knowledge with everyone.
- Feasible traffic and sharing

Use case model:

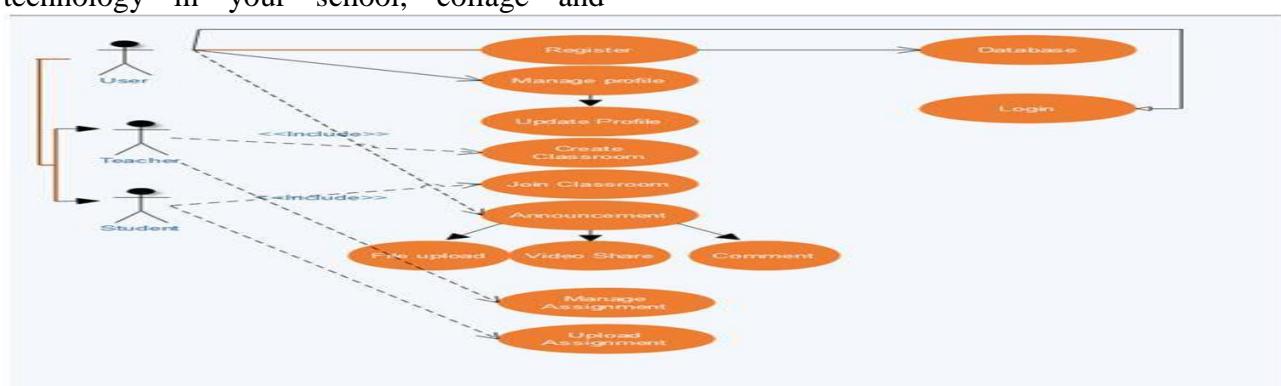


Fig. 1 Use case model of the proposed system

All user need to register first for login any time anywhere. Teacher can create one or many classroom with their own choice. Then trainer can glance a blank classroom where they can't see any post. They only can see some basic information that subject name, class code, section. Teacher requirement to provide that class code with their student.

Student can join in theirs classroom with class code. They can see all activity in theirs classroom. Teacher and student can give a post, comment, share video and many more. Teacher can manage course materials, assignment and many more.

6. DATA MODEL REQUIREMENTS

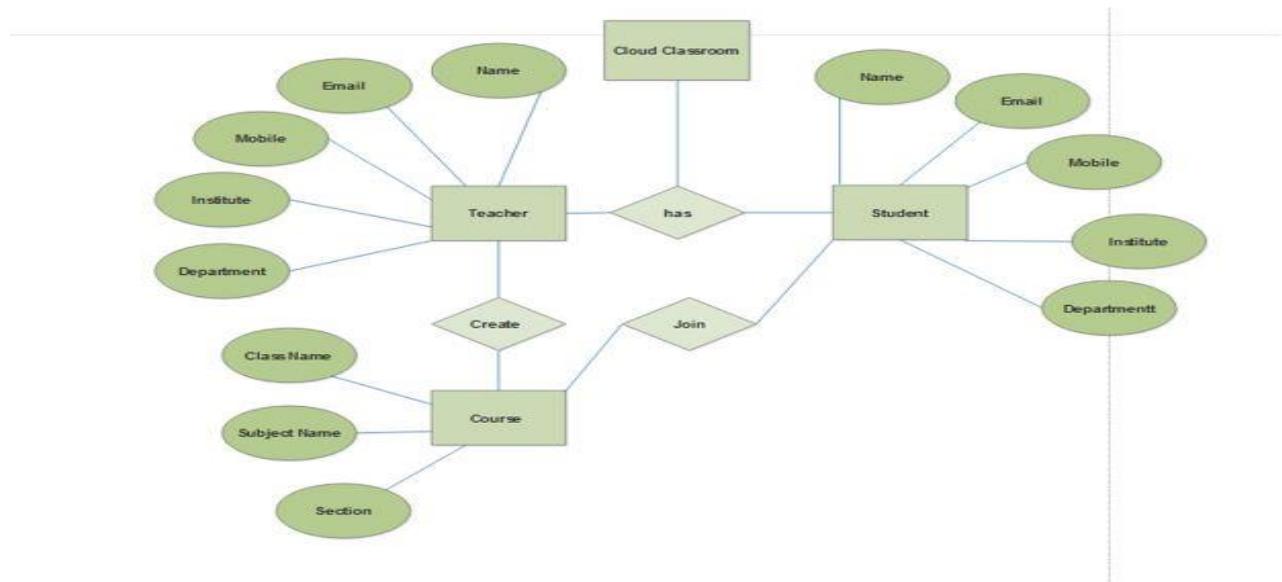


Fig: 2 Data model of proposed system

The entity relevance sketch and data pattern is emerged on The appreciation of real world that structure of an ingathering of Intention called essence, and of relevance between these Essence. The data pattern of the raised method is exhibited in The Fig. 2.0. Emerged on the metadata requirements, the database Blueprint is exhibited in the following Fig. 3.0. Requirements investigation used the way of sympathetic of The user emergent and expectation from the mentioned method or Entreaty. Requirements are to advantage defined of how the Raised method should act. The software necessity Exploration method also tunicate the intricate ought of invent and Documenting the requirements of whole these users, modelling and As a basis for method sketch.

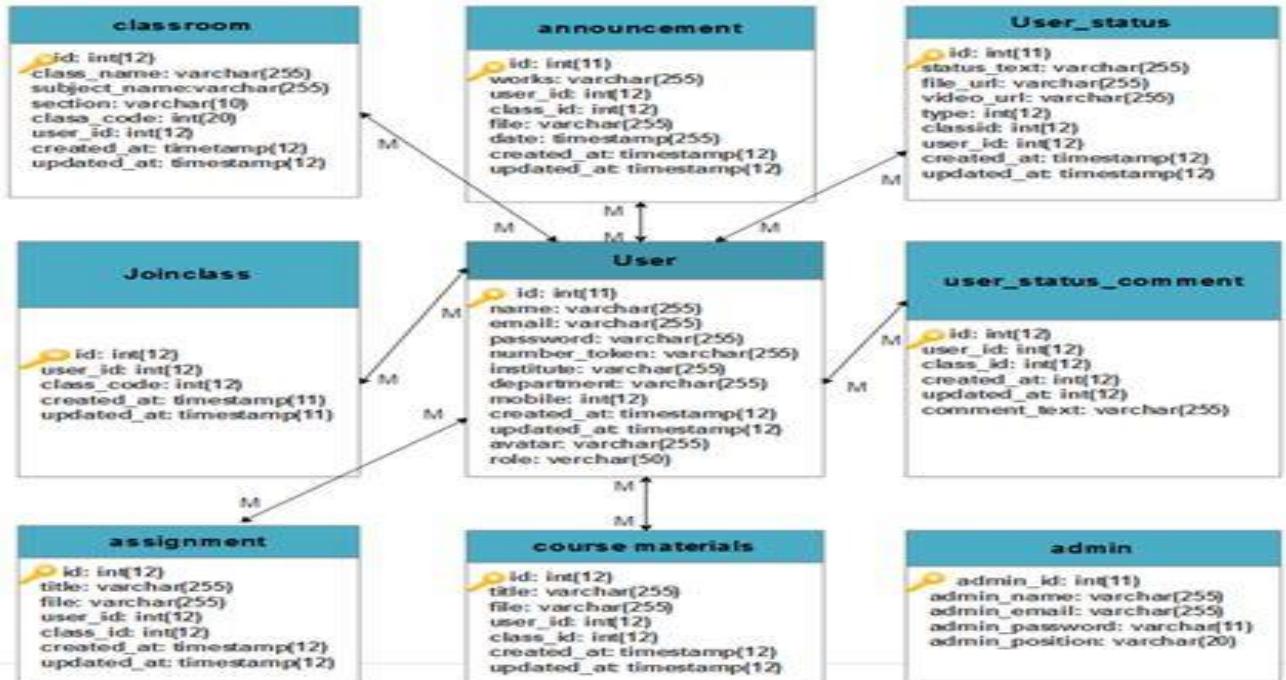


Fig: 3 Schema of the proposed system

7. IMPLEMENTATION AND TESTING

The proposed model is implemented using PHP object model and an amply informed MVC framework called Laravel.



Fig: 4 Home page of the proposed system

As exhibited in the Fig. 4, the home page contains different elements of the proposed system which includes both the some features

and components. The user, can know about this classroom throw this home page.



Fig: 5 Classroom of the proposed system

Suppose Mr. Shah is a teacher of system analysis and design. He need to originate a classroom for his learners. He need to register as a teacher to define his role teacher. And his student also concernment to define as a student in register page. After register teacher originate a classroom by clicking create classroom.

After creating a classroom by the teacher he get one class code in his classroom. Then he need to provide that class code to his students. After getting the class code student can join that class by putting that class code in join a classroom option.

Suppose teacher has been originate a classroom which name is System Analysis & Design? He need to upload a very momentous annunciation for his student that "Dear student I can't able to take your class tomorrow. So have fun and see you next class".

Now if teacher need to upload his class note by doc, pdf file he can upload easily in his classroom and student can see, read, view and download also that file. Suppose teacher need to upload a class note for system analysis class Lecture 1.

There are many YouTube tutorial for any subject related. Now if teacher need to percentage any subject related YouTube video tutorial he can do it easily. Suppose System

Analysis and Design teacher need to share a tutorial from YouTube then he simply go to the video URL and copy the URL id and pest video share option and suitable title is needed.

Teacher can post an announcement, upload a lecture and share a video. If students need to discuss something about the topic with the teacher then they execute comment anything in using comment box. In comment box teacher can provide answer for student.

System analysis and design is a core course of computer science and engineering. Numerous students take out this course. Teacher imperatives to give a work topic and take out the work from his students.

If teacher give an assignment topic with a due time and take out the assignments from his students with hardcopy it will be a long process. To unravel the long process virtual classroom will be a great solution. Let the assignment topic is "Differences between Red Box Testing and White Box Testing".

Then student can closed their work in virtual classroom. Students work visible only for the teachers. Student can't review or edit the works after uploading.

Suppose for the system analysis and design class teacher wants to upload all course

materials by zip folder for his students. Yes he can upload all course materials by clicking on course materials option. He must need to upload all files by zip folder.

In the virtual classroom teacher can see and counselor the student. Student can also become acquainted theirs classmates.

8. STUDENT FEEDBACK

A metering was driven from the students of one section (about 50 students) of student

sagacity of using virtual classroom. They leave some comments after using virtual classroom.

Some comments are here:

- Virtual classroom gives us a location to interlude.
- I think it's save lot of time.
- Virtual classroom is so user friendly.
- We can discuss easily via comment box.

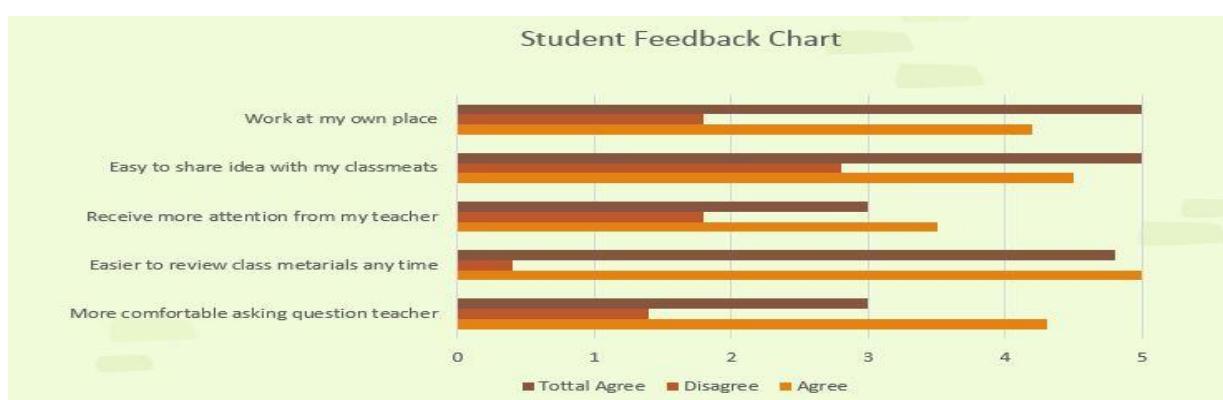


Fig: Student feedback in proposed system

9. LIMITATION

The virtual classroom has some limitations. It does concernment computer, internet and at the very first Levin. It has another limitation, it may accommodate the user to generate use of plagiarism. The mathematical way of grading the virtual student may need some more advancement as well.

According to Shibli Shahriar, the following are the hindrances in engaging students in their university for their learning feedback system (LFS) [11]:

1. Fear of mistakes: Students might be afraid of mistakes when they want to join in the discussion with other fellow students.
2. Access to Internet: Some students don't have any access to internet or they may think that using internet will be costly for them compared to its benefits derived from writing.

3. Lack of creativity: Many students didn't learn to think as they feel teachers will give

Those questions and they will memorize them to enhance the participation and to reduce the dropout rate, the followings are suggested:

A tutorial class on practically using LFS.

Sufficient computers and internet facility.

Intrinsic and extrinsic motivation to participate in discussion Designing questions that make sure to stop copy and paste culture

10. Conclusion

Technology gives us a new measure in learning and teaching. In this time of technology teacher should teach their students to contribute the world. There is so much that students can do with the Internet. Not only can they communicate with international students, they can gain from

others' knowledge and experiences, participate in classrooms, share ideas and solutions and learn about the many diverse cultures out there.[12]. Tim O'Reilly & Battelle (2009) has favored Web 2.0 as a better and quicker adaptation of students to technology, to the extent that Internet has become a platform simple and easy to use that corresponds to their interests and personal needs benefiting the collective intelligence[13]. The virtual classroom scheme can be a scope to assign students in a learning conditions. Students can always be touch with their teachers.

REFERENCES

- [1] Classroom technologies narrow education gap in developing countries Steven Livingston Tuesday, August 23, 2016
- [2] Ben McNeely. Using Technology as a Learning Tool, Not Just the Cool New Thing, Educating the Net Generation February 2005
- [3] Subhenur Latif and Narayan Ranjan Chakraborty "Virtual Community in Interactive Teaching: Five Cases." e-ISSN: 2278-0661, p-ISSN: 2278-8727 Volume 15, Issue 6 (Nov. -Dec. 2013), PP 69-75
- [4] Yousuf Islam. "Tertiary Education in Bangladesh –Brief History, Problems and Prospects", International Journal for the Scholarship of Teaching and Learning, Vol. 5, No. 2, pp-4, July 2011.
- [5] Integrating Technology into the Classroom Jong H. Chung United States Military Academy, West Po in t, NY, 2007.
- [6] 2013.University of the Sciences Website. [Online]. Available:
<http://www.usciences.edu/teaching/tips/spal.shtml>
- [7] Integrate Technology In The Curriculum As An Effective Teaching Strategy Dr. Fatma Abdullah. Available:
<http://www.ijarsite.com/wp-content/uploads/2016/03/IJAR-4-1-2016.pdf>
- [8] Using the technology of today, in the classroom today. Eric Klopfer, Scot Osterweil, Jennifer Groff, Jason Haas. Available:
http://dmlcentral.net/wp-content/uploads/files/GamesSimsSocNets_EdArcade.pdf
- [9] Technology Integration |Ideas That Work. Available:
http://www.educationworld.com/a_tech/tech/tech176.shtml
- [10] 2013.University of the Sciences Website. [Online]. Available:
<http://www.usciences.edu/teaching/tips/spal.shtml>
- [11] Shibli Shahriar (2011).Daffodil International University Forum. Available:
<http://forum.daffodilvarsity.edu.bd/index.php?topic=4622.msg21529#msg21529>
- [12] Internet Has Many Benefits Available:
<http://iml.jou.ufl.edu/projects/STUDENTS/Lui/index3.htm>
- [13] O'REILLY, Tim & BATTELLE, John; (2009). Web Squared: Web 2.0 Five Years On. Available:
<http://conferences.oreilly.com/web2summit/web2009/public/schedule/detail/10194>

Self Organizing Migration Algorithm with Curvelet Based Non Local Means Method For the Removal of Different Types of Noise

Sanjeev K Sharma

Associate Professor, Department of E&I
SATI, Vidisha (M.P.)
san0131966@gmail.com

Dr. Yogendra Kumar Jain

Professor and I/C HOD, Department of CSE
SATI, Vidisha (M.P.)
ykjain_p@yahoo.co.in

Abstract—This research focus on image sharpness and quality using a self-organizing migration algorithm (SOMA) with curvelet based nonlocal means (CNLM) denoising is presented. In this paper, first transform curvelet is used on the noisy image obtain image. Find the comparison of 2 pixels in the noisy picture which is evaluated depend on these curvelet produced pictures which include complementary picture capabilities at particularly excessive noise levels and the noisy picture at especially low noise levels. Then pixel comparison and noisy photograph are used to denoised end outcome found applying NLM technique. SOMA obtains better quality with the aid of varying threshold on the basis of image pixels. The threshold can be determined using lower and upper value of noisy image. Quantitative evaluations illustrate that the proposed scheme perform more enhanced than the other filters namely median filter (MF) progressive switching median filter (PSMF), NLM, CNLM denoising process in conditions of noise removal and detail protection. Using different parameters for example Peak Signal Noise Ratio (PSNR), means Structural Similarity Matrix (MSSIM) and SSIM for noise free image. It is illustrated that the improved scheme provides an excessive degree of noise removal whilst maintaining the edges and other information in the image. In this study, algorithm is tested on dissimilar kind of noise explicitly, Random Valued Impulse Noise (RVIN), Gaussian Noise and Salt and Pepper (SNP) Noise with varying noise density from 10 to 90%. The proposed system proves better performance on high noise density.

Keywords— SOMA, Impulse Noise, CNLM, PSNR, MSSIM, SSIM, PSMF, Median Filter, Gaussian Noise, Salt and Pepper Noise.

I. INTRODUCTION

Denoising is the term of recapture version of data pixels from the noisy photograph model via smoothing it out with admire to its surrounding pixels. In photo processing (IP) that is very vital preprocessing step before these pictures are analyzed. Hyper spectral Imagery belongs to the faraway sensing region in which these pictures are remotely sensed with committed sensors. These sensors are designed to discover a worldwide distribution of object models from the target region, it captured or accumulated information or photos and provide to programs wherein those photo are studied [1].

In general, the obtained image is mostly useless with many types of noise or degradations or noise when the imaginary is

generated or in the process of transmission. Thus the corrupted image has necessity to be processed before they are used in some real applications. Those inverse problems contain image reconstruction, image restoration and image denoising [2].

II. NOISE MODEL

Type of Noise:

A. Impulse Noise

The model of impulse noise comprises two different impulse values with probability which is equal. These are the least and most pixel values of the taken into consideration integer c program language period (i.e., 0 and 255 for an 8-bit picture). The minimum pixel value, i.e., a black pixel is called poor impulse or salt and the maximum pixel price, i.e., a white pixel is known as high-quality impulse or pepper [3]. In RVIN noise is spread consistently. Dynamic range [0, 255] may take by RVIN. Previous to bring in the proposed framework, we first outline the 2 most generally applied impulse noise fashions used in this paper.

B. Gaussian Noise

It is uniformly dispersed over signal [Um98]. Here, in noisy image every pixel is combination of random Gaussian distributed degradations value and pixel value.

C. Salt and Pepper Noise

It has most effective two unique viable values. The each chance is classically not up to 0.1. It is kind of noise which is obvious in image. It represents as white and black pixels. Strong degradations discount system morphological or a median filter. The salt and pepper degradations are customarily prompted through pixel elements malfunctioning within the digital camera sensors. [6].

III. DENOISING FILTERS

3.1. Progressive Switching Median (PSM) Filter Method

A novel median-based switching clear out, called PSM filter, on this both noise clear out and impulse detector are using progressively in the iterative manner. A main benefit of such a technique is that some impulse pixels placed in large noise

blotches center also can be successfully detected and filtered.[22]

3.2. Median Filter Method

Median value is the worth in the center role of any taken care of series [7].

Image de-noising manner founded some median filter (MF) were proposed, some MF are software program oriented. Some of the large processes were explained underneath.

3.3. Simple Non Local Means Method (SNLM)

Mostly, the NL-method approach approximation an innocent depth as not unusual weighted for all pixel inside the picture, and the weighted proportional value. This technique is also to do away with aggregate of RVIN, SNP and GN. The nice answer may be to domestically various parameters, so that they're primly tuned to do away with the precise amount and diverse noises found in each part of the photograph [22].

3.4. Curvelet Based Non Local Means Method (CNLM)

The key to the CNLM technique dishonesty in similarity weight estimate. To describe the way to outline pixel similarity, we are able to take a look at the reconstructed picture corresponding characteristics to every curvelet scale. Provide a degradations image L_0 ; we 1st decompose it into n-scale curvelet coefficients (CC) applying Eqs. (7)–(13)(n value is define eith the aid of picture width [23]).

Considering which curvelet transform (CT) maps the photograph noise into specific scales in the frequency area to gain rather little coefficients, it is simple to apprehend which the noise at the recon-strutted picture at every degree could be significantly attenuated in comparison with that in noisy picture. Therefore, those reconstructed images match weight founded on compute can make easy sup-urgent the noise have an effect on disadvantageous.

3.5. Self-Organizing Migration Algorithm (SOMA)

SOMA is depending on self-organizing person group's conduct in "social environment". Only location of the individuals within the investigated area is modified in the course of era called "migration loop". The algorithm, evolved through approach of prof. Zelinka in 1999. Numerous dissimilar description of SOMA exists. All primary each-to-One SOMA objectives essential for accurate information of the set of rules are explain beneath:

Parameter definition: Before beginning the set of rules, SOMA's parameters Step, Path Length, PopSize, PRT and the Cost Function required to be defined. The Cost Function is really a function which returns a scalar that may right away function a level of fitness.

Creation of Population: Population of humans is generated randomly. Each parameter for every separate ought to be selected randomly from the given range.

Migration loop: All people from populace (PopSize) is predicted via the Cost Function and the Leader (person with the best condition) is selected for the prevailing migration loop. Then all different people begin to bounce, (consistent with Step limitation description) in the course of the Leader. All individual is estimated in the end bounce applying the Cost Function. Jumping continues till a novel function described through Path Length has been reached [9].

3.6. Stein's Unbiased risk Estimator

Wavelet is a Multi-resolution Analysis (MRA) process eliminating data capable from an image (or a signal) space-frequency resolutions varying. If achieve a gray picture wavelet decomposition, because of wavelet representation sparsity, signal component is classically concentrated in a few high amplitude coefficients even as noise is spread uniformly throughout all coefficients. This bureaucracy the premise of wavelet shrinkage denoising. If we decrease to zero wavelet coefficients having amplitude less than a elect threshold value, most of the noise would be eliminated from the picture.

The main steps of wavelet shrinkage are:

1. Computing DWT of the original picture
2. Thresholding of wavelet coefficients.
3. Performing IDWT

IV. PROBLEM DEFINITION AND OBJECTIVE OF PROPOSED WORK

In previous curvelet based non local means (CNLM) method has many issues for removing high density noisy pixels. CNLM method removes noise on low density, but for the high density noise it did not work properly and image quality also degraded. And quality measures did not achieve good results in terms of PSNR, MSSIM and SSIM. For resolving these factors, suggest an evolutionary algorithm with an CNLM method for improving previous results. This proposed algorithm determined optimal set of parameters and refine the results. But the sometime previous method gives better result as equated to propose because of learning algorithm. It detects noise on small window size easily with high density noise.

V. PROPOSED METHODOLOGY

In the proposed work, implement the combination of SOMA and CNLM, the noisy pixels are locating in a pretty narrow range and therefore can decrease the opportunity of wrong finding. In the proposed system, the photograph to be denoised is divided into sub-imaginary. For any given $M \times M$ gray level image that is defined thru: $M \times M \rightarrow I$ where $I = [r, c]$ stand for the variety of pixel values. Pixel value at position (i, j) is given thru $F(i, j)$.

In the process of denoising requires selection of the kind of wavelet basis function (mother wavelet) to be used, the phase of decomposition and the threshold value for each level of decomposition. In this paper, we employ SOMA to find an optimal value for below variables. To find the optimal solution, we take threshold on the lower and upper value.

- The parameters which govern the convergence and performance SOMA behavior are: [11]
- Various individuals and their dimension
 - Various iterations (migration loops)
 - Path Length: It position governs at which an separate will stop while leader following.
 - Step Size: It decides the granularity of the path towards the leader.
 - PRT: Pattern created applying this variable directs the

TABLE I. SET VARIABLES FOR OPTIMIZATION[12]

Variables to be optimized	Permitted Values
Types of Wavelet	Daubechies (db4, db6, db8), Symlet (sym4, sym6, sym8, sym10), Coiflet (coif2, coif4)
Decomposition Level	1-4
Threshold	It is estimated using lower and upper threshold

Proposed Algorithm

1. Consider 'I (x)' as the depth value of picture.
2. Add RVIN pixels into input picture, I at pixel position x and $[d_{\max}, d_{\min}]$ the dynamic range of I. Dynamic range of gray levels image. For 8-bit pictures, $d_{\max} = 0$ and $d_{\min} = 255$. [4]

$$I_{\text{noisy}}(x) = \begin{cases} d_x & \text{with probability } r \\ I(X) & \text{with probability } (r-1) \end{cases} \quad (1)$$

Here d_x is uniformly distributed in $[d_{\max}, d_{\min}]$ and r defines the random-valued impulse noise level [5].

3. Add Gaussian distribution, probability distribution function is shaped of bell,

$$F(g) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(g-m)^2/2\sigma^2} \quad (2)$$

Where g (gray level), m (mean) or standard functions and σ is (standard deviation (SD)) of the noise. [6]

4. Consider that the gray degrees of any pixel value, in any window (wx) of size n Xn are represented by $X_1, X_2, X_3, X_4, \dots, \dots, X_1$ and it becomes $X_{i1} \geq X_{i2} \geq X_{i3} \geq \dots, \dots, X_{in}$ after sorting it in descending or in an ascending order

$$M_X = \text{Median}(W_X) = \begin{cases} X_{i(n+1)/2}; & n \text{ is odd} \\ \frac{1}{2} \left[X_{i(\frac{n}{2})} + X_{i(\frac{n}{2})+1} \right]; & n \text{ is even} \end{cases} \quad (3)$$

5. Let fX f be length of the quest window “ Ω ”, shaped through partitioning of an photograph. A general filtering window “ Ω ”, is given in Eq. (4), it has rxr

matrix. The gray stage at any pixel (i,j) is stand for with the aid of $I_{(i,j)}$

$$\Omega = \begin{bmatrix} I_{1,1} & I_{(1,\frac{f+1}{2})} & I_{1,f} \\ I_{(\frac{f+1}{2},1)} & I_{(\frac{f+1}{2},\frac{f+1}{2})} & I_{(\frac{f+1}{2},f)} \\ I_{f,1} & I_{(f,\frac{f+1}{2})} & I_{f,f} \end{bmatrix} \quad (4)$$

6. NL- means approach estimation an innocent depth as commonplace weighted for all pixel in the photograph, and the weighted proportional value

$$NLM(i) = \frac{1}{C(i)} \sum_{j=\Omega} w(i,j) L_0(j) \quad (5)$$

Wherein Ω is the quest window and $C(i) = \sum_{j=\Omega} w(i,j)$ is a ordinary-ization constant; the weight $w(i,j)$ indicates the matches among picture patches N(i) and N(j) (i.e., comparison windows) targeted at 2 pixels i and j.

7. Then, the reconstructed pictures are acquired the use of booking the CC at every scale and location the coefficients at closing scales to 0. Let $L_q (1 \leq q \leq n)$ signify the reconstructed picture at qth level using the curvelet coefficients on the qth scale. The match weight amid pixels i and j inside the photo $L_m (0 \leq m \leq n)$ might be defined as:

$$W_m(i,j) = \exp \left(-\frac{\|L_m(N(i))-L_m(N(j))\|_2^2}{h_m^2} \right), \quad (6)$$

where $h_m = \left(\frac{\text{med}(HH)}{0.6745} \right)$

Where h_m is a constant comparative to the noise SD the image L_m which can be expected by the process used. $\text{med}(HH)$ Is wavelet coefficient (Low pass filter and High pass filter).

$$\sum_{j=-\infty}^{\infty} w^2(2^j r) = 1 \quad r \in \left(\frac{3}{4}, \frac{3}{2} \right), \quad (7)$$

$$\sum_{l=-\infty}^{\infty} v^2(t-l) = 1 \quad r \in \left(-\frac{1}{2}, \frac{1}{2} \right), \quad (8)$$

$$U_j(r, \theta) = 2^{-\frac{3j}{4}} W(2^{-j} r) V \left(\frac{\frac{1j}{2}}{2\pi} \right) \quad (9)$$

Where $ij/21$ is the integer part of $j/2$ W and V restriction the aid U_j to a polar wedge that is symmetric with appreciate to foundation. classify the waveform $\phi_j(x)$ thru means of its FT $\phi_j(w) = U_j(w)$. If equispaced rotation angles sequence $\theta_1 = 2\pi, 2^{\frac{j}{2}}$, $l (0 \leq l \leq 2\pi)$, and order of translation parameters $k = (k_1, k_2) \in Z^2$ are introduced, the family of curvelets $\phi_{j,l,k}$ will be defined at scale 2^j , orientation θ_1 and position $x_k^{(j,l)} = R_{\theta_1}^{-1}(k_1 \cdot 2^{-j}, k_2 \cdot 2^{-j/2})$ as:

$$\phi_{j,l,k} = \phi_j(R_{\theta_1}(x - x_k^{(i,j)})) \quad (10)$$

Where R_θ is the rotation matrix thru θ radians and R_θ^{-1} is its inverse defined as:

$$R_\theta = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \quad (11)$$

8. The following difficulty integral make the CT of a function $f \in L_2(\mathbb{R}^2)$:

$$c_{j,l,k} = (f, \phi_{j,l,k}) = \int f(x) \phi_{j,l,k}(x) dx \quad (12)$$

9. The coefficients $c_{j,l,k}$ of the equation are understand like the decomposition into a bases of curvelet features $\phi_{j,l,k}$ [17].

$$\delta = \min(t, h_m \sqrt{2 \times \log(\text{no. of pixels})}) \quad (13)$$

Where, δ is estimated threshold, t is the threshold which is estimated using upper and lower value that minimizes Stein's unbiased hazard estimator and h_m is the SD of noise, which is estimated using Eqn 13 coefficients at 4th decomposition level). [10]

Finalize the parameters listed above in table1

10. Initialize the population and calculate all individual fitness.
11. Choose leader (individual with maximum fitness).
12. Create PRT vector with the aid of equation (14)

$$PRT_{vec}(i) = 1; \text{ if } r \text{ and } < PRT \text{ for each dimension} \quad (14)$$

13. Update the position of individuals using equation (15)

$$x_p^{l+1} \leftarrow x_p^1 + (x_{leader}^{l+1} - x_{p,start}^{l+1}) \times step_size \times prtvec \quad (15)$$

14. Update fitness value and select new leader
15. Repeat steps 12 to 14 till a extinction condition is satisfied

VI. EXISTING TECHNIQUES

It presented that Dual Tree Complex Wavelet Transform (DT-CWT) with GCV is used which give ideal reconstruction over the conventional WT. Estimation is carried out in many parameters terms for example PSNR, mean Structural Similarity and Correlation Coefficient. The main weakness of this technique is better angular resolution which is not providing by DT-CWT and to overcome this problem 2D DTCWT method [13].

It supplied that growth denoising set of rules overall performance for commercial RT picture, an optimized wavelet denoising set of rules making use of hybrid noise model.

Smearing problem is main wavelet denoising algorithm problem [14].

It offered that an set of rules rely on WT and wiener clear out the use of log electricity distribution to denoised DI corrupted with the aid of Poisson-Gaussian noise. Poisson-Gaussian noise face some problem when a low level signal is expected, is very limited. Among the few contributions dealing with this problem [15].

It supplied that a singular way is proposed to dispose of GN found in fingerprint picture the usage of Stationary Wavelet Transform, a threshold founded on Golden Ratio and weighted median. A disadvantage is a completely massive redundancy and elevated computational complexity. The lack of directionality and oscillating persist because the stationary wavelet remodel is relying upon a filter out bank structure [16].

It presented that IR imaginary most have vague details and low resolution, resulting in poor visual effect and lower image quality. One weakness of K-SVD is that at the same time as doing well in reconstruction, it lacks discrimination functionality to separate special lessons. K-SVD requires massive garage because the computed non-0 coefficients reside in specific locations [17].

It provided that Non-neighborhood Means filters and diffusion tensor technique combination in 3D image denoising region. Non-local means algorithm is improved by taking symmetry advantage in weights and through applying a lookup table to speed up the weight computations. Diffusion pictures are sensitive to water diffusion that is in the 5-10 μm order at the time of size time. If this happens, images are sometimes full of ghosting because of the water molecules encountering obstacles. Because of this, DTI need at least 7 tensor fittings, requires wide computing power, man-hours, and expertise [18].

It presented that a novel imaginary denoising way depend on the mixture of SWT and bilateral filter. The essential giving of this paper is inside the utilization of a brand new neighborhood association to broaden a brand new multiscale bilateral clear out. The some trouble in this paper first is Gradient reversal - creation of false edges within the imaginary and second is Staircase impact - intensity plateaus that result in imaginary performing like cartoons [19].

It presented that denoising earlier demosaicking strategy is exploited. Some problem find out in this paper Signal and noise must both be random, numerous programs have a deterministic signal and random noise and Extend Wiener clear out (or Phillips method) to allow deterministic signal [20].

It presented that the algorithm depend on stack to eliminate the small vicinity noises in binary microalgae picture. The

denoising of picture is accomplished via using this set of regulations via scanning picture one time. The new algorithm now not handiest conserves sign's authentic capabilities; however additionally has stronger capacity to cast off noise. The numerical experiments have illustrate that the set of rules is very powerful in noise discount, and is higher than conventional approach in complexity and running time of the image denoising. This technique better result provide binary image not for another varieties of image [21].

VII. COMPARATIVE TABLES

The experimental outcomes have been implementing using MATLAB12 on Image Processing. The outcomes have been experienced on gray scale with dissimilar varieties of format images of size 256 X256. The evaluation of the proposed method is obtained on the basis of PSNR, SSIM and MSSIM. We have implemented image denoising on various images corrupted with RVIN, Gaussian noise (GN) and pepper and salt noise on different noise density. It is varying from 10% to 90%. The denoised images are evaluated using three criteria mentioned below: For the de-noised image "Z", of size M X M, the PSNR is given by:

- A. *PSNR*: It is exploited to measure the visible best of the denoised photograph in comparison to the particular picture. Compute PSNR and MSE value of a denoised and unique image.

$$MSE(x) = \frac{1}{M} \|I - Z\|^2 = \frac{1}{M} \sum_{i=1}^M (I - Z)^2 \quad (16)$$

Where I is the original image

$$PSNR(x) = \frac{10 \log ((double(m).^2))}{MSE(x)} \quad (17)$$

Where m is the original picture maximum value.

- B. Structural Similarity Matrix (SSIM)- It is used for calculating similarity content between original image and denoised image.

$$SSIM = \frac{(2\mu_x\mu_y - c_1)(2\sigma_{xy} - c_2)}{(\mu_x^2 + \mu_y^2 - c_1)(\sigma_x^2 + \sigma_y^2 - c_2)} \quad (18)$$

Where μ_x is the average of x, μ_y is the average of y, σ_{xy} is the covariance of x and y, $c_1 = (K_1 L)^2$, $c_2 = (K_2 L)^2$, $K_1 = 0.01$ and $K_2 = 0.03$ by default and L is the dynamic variety of pixel values.

- C. MSSIM-

$$MSSIM(E, F) = \frac{\sum_{k=1}^R SSIM(x, y)}{R} \quad (19)$$

Where R is the entire number of local windows in the picture, E and F signify the unique picture and the denoised image, respectively; x and y are the picture contents at the k-th local window in the real and denoised pictures

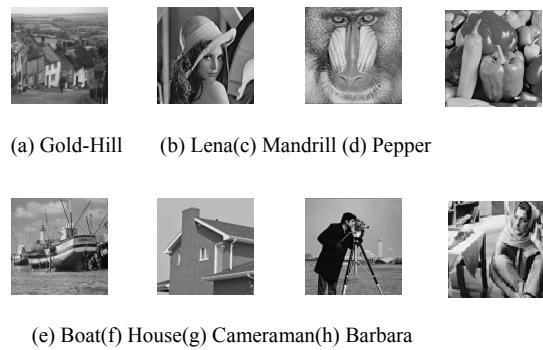


Fig. 1. Grayscale test original images of 8-bit per pixel.

TABLE II. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES BY GAUSSIAN NOISE USING $\sigma = 40$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 40$	Pepper $\sigma = 40$	Lena $\sigma = 40$
MED[7]	13.27/0.891	13.79/ 0.881	15.33/0.893
PSMF[8]	12.70/ 0.873	13.15/ 0.860	15.31/0.902
NLM[21]	6.91/ 0.592	7.34/ 0.592	8.97/0.597
CNLM[23]	17.98/ 0.977	18.88/ 0.972	19.85/0.982
Proposed	37.49/1.004	38.50/1.003	38.21/1.006

TABLE III. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES BY GAUSSIAN NOISE USING $\sigma = 70$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 70$	Pepper $\sigma = 70$	Lena $\sigma = 70$
MED	13.54/0.890	14.03/0.892	15.67/0.899
PSMF	12.94/0.871	13.35/0.872	15.60/0.903
NLM	7.20/0.605	7.64/0.607	9.303/0.610
CNLM	18.06/0.975	18.76/0.972	20.26/0.984
Proposed	37.68/1.004	38.71/1.004	37.84/1.006

TABLE IV. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES BY GAUSSIAN NOISE USING $\sigma = 90$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 90$	Pepper $\sigma = 90$	Lena $\sigma = 90$
MED	13.73/0.902	14.32/0.901	15.75/0.899
PSMF	13.10/0.882	13.61/0.881	15.66/0.903
NLM	7.43/0.623	7.78/0.617	9.46/0.618
CNLM	18.37/0.983	19.30/0.982	20.36/0.977
Proposed	37.57/1.004	38.64/1.003	37.99/1.006

TABLE V. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES BY RVIN NOISE USING $\sigma = 40$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 40$	Pepper $\sigma = 40$	Lena $\sigma = 40$
MED	22.62/1.018	22.96/1.011	21.97/0.971
PSMF	23.04/1.007	23.33/1.004	22.71/0.977
NLM	14.28/0.974	13.85/0.931	13.40/0.790
CNLM	18.74/1.074	18.42/1.043	17.57/0.906
Proposed	37.54/1.003	39.28/1.001	38.21/1.003

TABLE VI. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES BY RVIN NOISE USING $\sigma = 70$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 70$	Pepper $\sigma = 70$	Lena $\sigma = 70$
MED	16.64/1.039	15.89/1.018	15.00/0.839
PSMF	16.85/1.018	16.21/1.009	15.34/0.852
NLM	11.68/0.939	11.20/0.883	10.67/0.660
CNLM	16.84/1.122	15.81/1.071	14.75/0.822
Proposed	37.67/1.003	38.63/1.003	38.07/1.005

TABLE VII. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES BY RVIN NOISE USING $\sigma = 90$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 90$	Pepper $\sigma = 90$	Lena $\sigma = 90$
MED	13.78/0.897	12.96/0.991	12.00/0.722
PSMF	13.13/0.878	12.96/0.986	12.03/0.725
NLM	7.38/0.616	9.98/0.837	9.38/0.594
CNLM	18.44/0.977	14.16/1.050	12.93/0.753
Proposed	37.37/1.004	38.71/1.003	37.81/1.006

TABLE VIII. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES USING $\sigma = 40$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 40$	Pepper $\sigma = 40$	Lena $\sigma = 40$
MED	18.07/0.980	18.56/0.980	18.04/0.941
PSMF	22.14/1.002	22.97/1.003	22.98/1.004
NLM	9.99/0.811	9.72/0.768	9.40/0.605
CNLM	18.50/1.076	18.09/1.053	17.35/0.901
Proposed	37.57/1.003	38.83/1.003	37.79/1.005

TABLE IX. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES USING $\sigma = 70$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 70$	Pepper $\sigma = 70$	Lena $\sigma = 70$
MED	9.89/0.818	9.89/0.765	9.67/0.624
PSMF	9.86/0.815	9.86/0.760	9.633/0.622
NLM	7.26/0.678	7.09/0.623	6.90/0.461
CNLM	16.15/1.118	15.36/1.046	14.42/0.809
Proposed	37.74/1.004	38.78/1.003	38.18/1.006

TABLE X. DIFFERENT TECHNIQUES EVALUATE PSNR (DB) AND MSSIM OVER THREE DEGRADED IMAGES USING $\sigma = 90$ (WHERE PSNR VALUE IS THE LEFT OF '/')

Method	Boat $\sigma = 90$	Pepper $\sigma = 90$	Lena $\sigma = 90$
MED	6.63/0.653	6.57/0.606	6.37/0.433
PSMF	6.62/0.652	6.56/0.603	6.36/0.432
NLM	6.02/0.610	5.85/0.560	5.67/0.391
CNLM	14.73/1.128	13.72/1.044	12.69/0.741
Proposed	37.65/1.004	38.79/1.004	37.87/1.005

VIII. RESULT ANALYSIS

Fig. 2 indicates the PSNR and MSSIM values for the compared filters in use on the despoiled Boat picture on Gaussian Noise (40% noise density) and proposed filter out proved higher than different filters. Tables 2 to Table four listing the PSNR and MSSIM values of all the estimate strategies operating on pix Boat, Pepper and Lena with $\sigma = 40$ to $\sigma = 90$, the use of Gaussian noise respectively. Obviously, the SOMACNLM filter proved better than other evaluated filters in terms of PSNR and MSSIM. Tables five to Table 7 list the PSNR and MSSIM values of all the evaluated procedures running on photographs Boat, Pepper and Lena with $\sigma = 40$ to $\sigma = 90$, the use of Impulse noise respectively. Tables 8 to Table 10 list the PSNR and MSSIM values of all the evaluated approaches in use on pictures Boat, Pepper and Lena with $\sigma = 40$ to $\sigma = 90$, using SNP noise respectively. When increases the noise density, then PSNR and MSSIM decreases for all filters, but proposed shown in Tables it gives improve value of PSNR and MSSIM at high density. As shown, proposed filter proved much high PSNR and MSSIM as equated to extra denoising filters. Outcomes obtained using various de-noising filters for Lena picture are illustrate in Fig 6 to Fig 9 on high density noise with all variety of noise and proposed filter shows better results on each noise and noise density. In Table 2, PSNR and MSSIM values acquired via extraordinary median filter and non nearby means primarily based de-noising process for Boat, Pepper and Lena picture are proven. As match up to to PSMF filtering founded-noising, proposed technique results in a progress of PSNR/MSSIM is varies from 37.49/1.004 to 38.21/1.006. The PSNR end result of the proposed technique is superior to each the another

approaches for noise densities up to 90%. The presentation of our proposed scheme, allowing for eight test pictures corrupted thru RVIN, GN and SNP noise is represented in Tables and Fig. 1 to Fig 9. Fig. 1 to Fig 9. In Fig. 2 to Fig nine, it's miles virtually seen that the presentation curve of our proposed manner is well higher than the prevailing de-noising filters. Between the existing schemes, CNLM filter shows the best results. On 3X3 window size reached the PSNR/MSSIM value high up to 90% noise density. The performance is dependent on image with add to in window size; number of pixels in any window will enhance which leads to better difficulty and time intake. For lower noise density, numerous pixels despoiled through noise are tons much less which ends up in excessive PSNR cost with lesser length of window. For maximum noise densities various pixels corrupted thru noise are substantial consequently the de-noised photograph PSNR with higher window length achieves higher fee because of higher accuracy on account of greater number of pixels underneath attention for finding of noise at any particular pixel. The higher noise density is removed by our proposed method because optimum solution. Image results several filters on Lena picture for 90% noise corruption through Gaussian Noise on 3X3 window are given in Fig. 7. PSNR/MSSIM cost of our proposed process is considered as 39.28/1.001dB, which is the nice among all. Further in Fig. Five the outcomes of Pepper picture are specified for 90% noise density on GN. Proposed scheme present the PSNR/MSSIM of 38.64/1.003 dB at this level. Figs. 8 and 9 show the outcomes of dissimilar filters in restoring Lena image tainted thru 90% noise density which is degraded using RVIN and SNP noise respectively. The restored picture first-rate of the proposed technique is the best, both quantitatively.

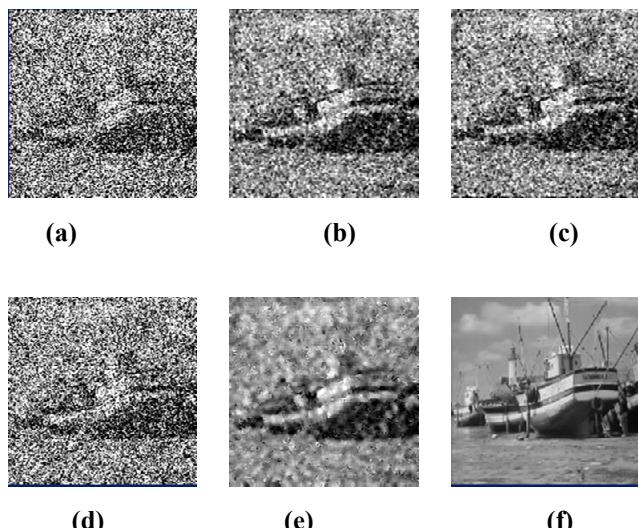


Fig. 2. Results of different filters in restoring Gaussian Noise 40% corrupted Boat image: (a) Noisy picture, (b) MED , (c) PSMF, (d) NLM, (e) CNLM (f) Proposed technique.

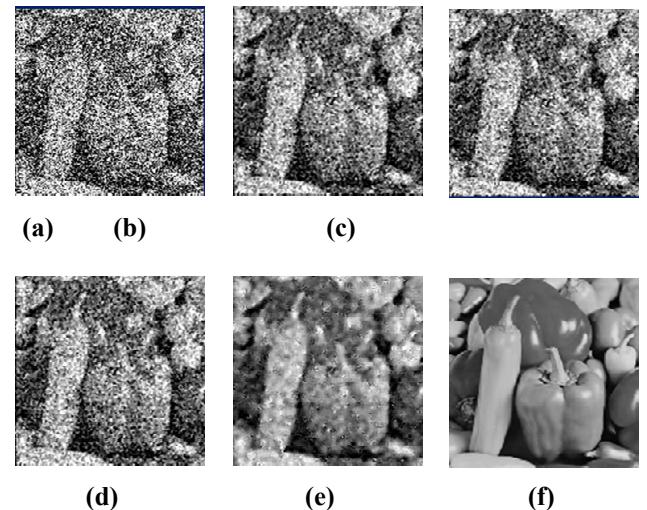


Fig. 3. Results of different filters in restoring Gaussian Noise 40% corrupted Pepper picture: (a) Noisy Image, (b) MED , (c) PSMF, (d) NLM, (e) CNLM (f) Proposed method.

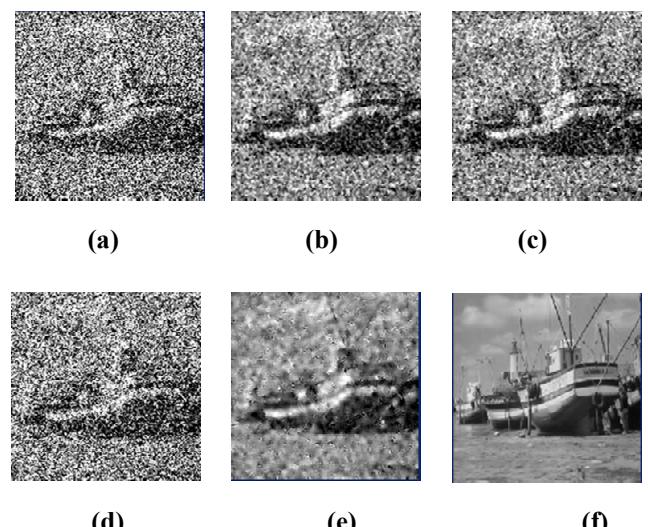


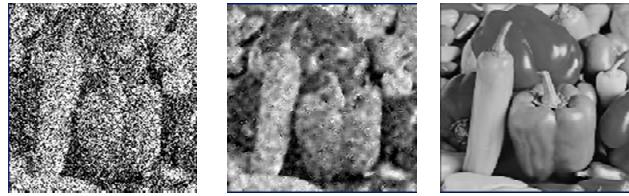
Fig. 4. Results of dissimilar filters in restoring Gaussian Noise 90% corrupted Lena image: (a) Noisy picture, (b) MED , (c) PSMF, (d) NLM , (e) CNLM (f) Proposed process.



(a)

(b)

(c)

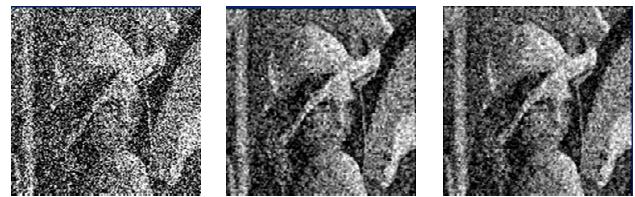


(d)

(e)

(f)

Fig. 5. Results of different filters in restoring Gaussian Noise 90% corrupted Pepper image: (a) Noisy Image, (b) MED , (c) PSMF, (d) NLM , (e) CNLM (f) Proposed method.



(a)

(b)

(c)



(d)

(e)

(f)

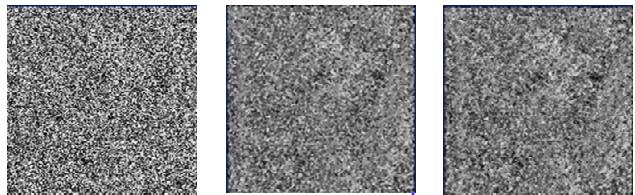
Fig. 7. Results of dissimilar filters in restoring Gaussian Noise 90% degraded Lena picture: (a) Noisy Image, (b) MED , (c) PSMF, (d) NLM , (e) CNLM (f) Proposed technique.



(a)

(b)

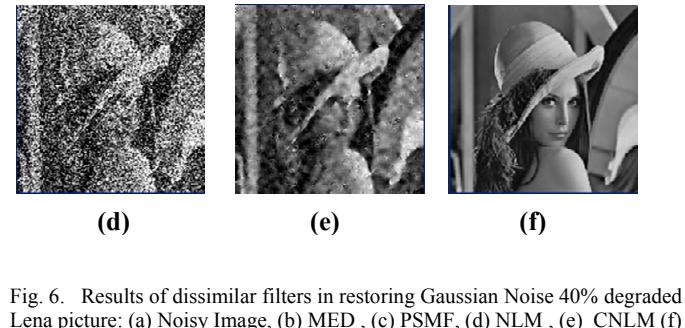
(c)



(d)

(e)

(f)

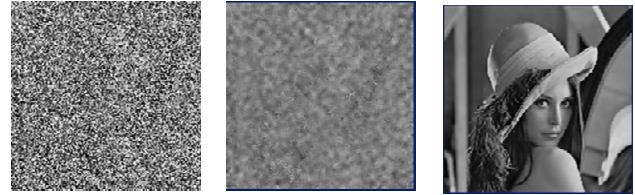


(d)

(e)

(f)

Fig. 6. Results of dissimilar filters in restoring Gaussian Noise 40% degraded Lena picture: (a) Noisy Image, (b) MED , (c) PSMF, (d) NLM , (e) CNLM (f) Proposed method.



(d)

(e)

(f)

Fig. 8. Results of dissimilar filters in restoring Impulse Noise 90% corrupted Lena image: (a) Noisy picture, (b) MED , (c) PSMF, (d) NLM , (e) CNLM (f) Proposed scheme.

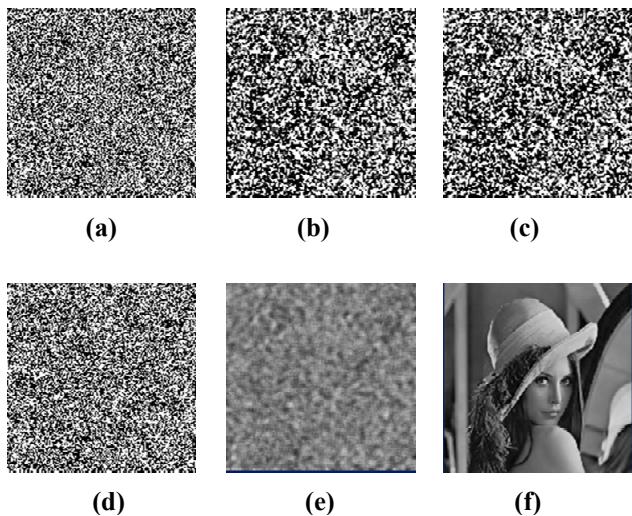


Fig. 9. Results of different filters in restoring SNP Noise 90% degraded Lena image: (a) Noisy picture, (b) MED , (c) PSMF, (d) NLM , (e) CNLM (f) Proposed method.

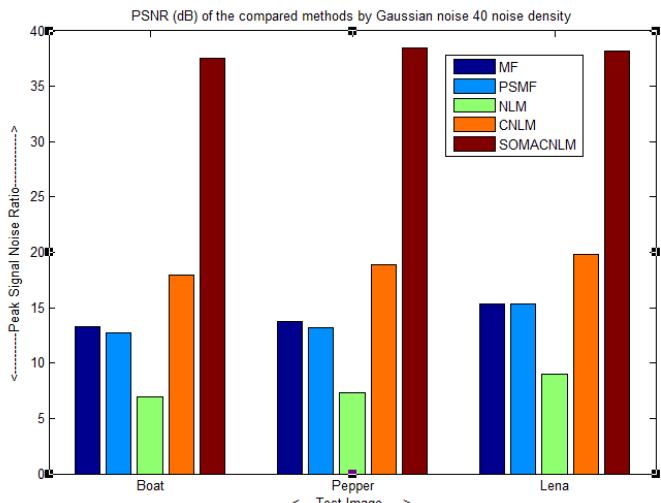


Fig. 10. Comparision chart of PSNR show of dissimilar filters specified in Table 1.

The observation from Fig. 10 shows that the CNLM methods can't suppress noise efficiently, the NLM method cause photograph over-smoothing near the edges and inside the textural regions, the CNLM process produces artifacts. By comparison, the SOMACNLM scheme offers the pleasant healing outcomes in that it may suppress noise efficiently even as pre-serving picture information regardless of in smooth areas or element areas.

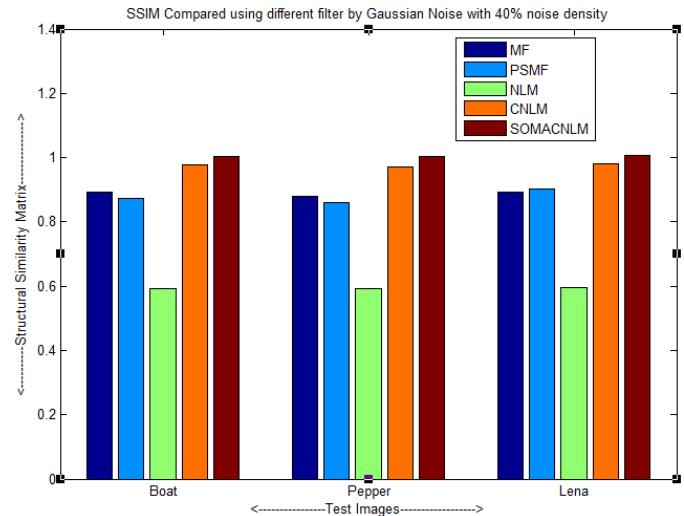


Fig. 11. Comparision chart of SSIM show of dissimilar filters specified in Table 1.

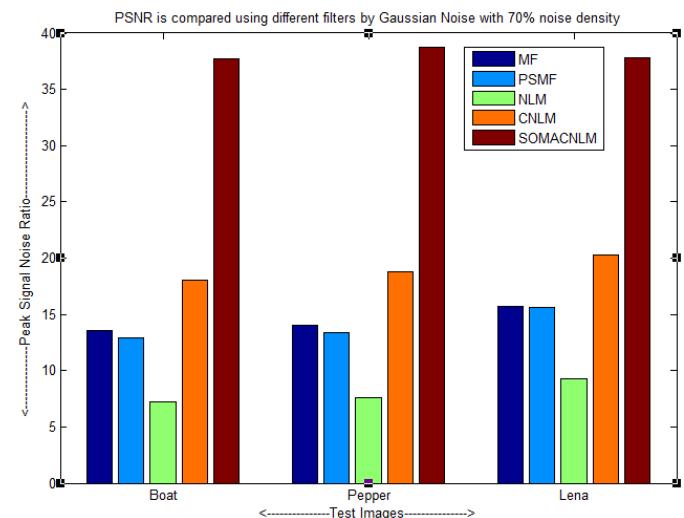


Fig. 12. Comparision chart of PSNR show of dissimilar filters specified in Table 2.

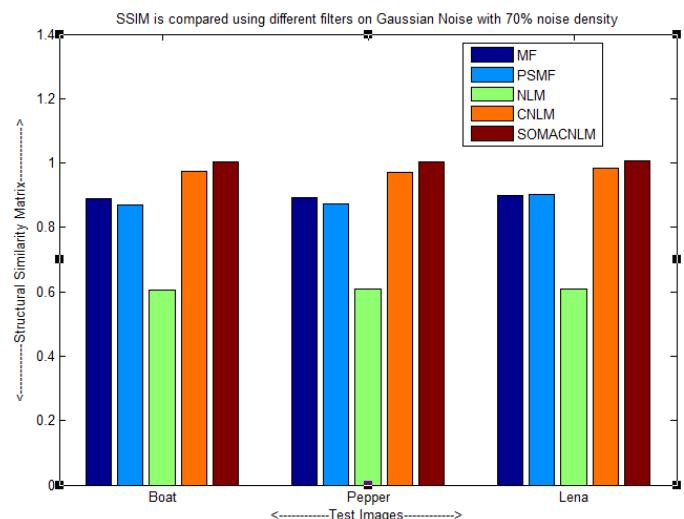


Fig. 13. Comparision chart of SSIM show of dissimilar filters specified in Table 2.

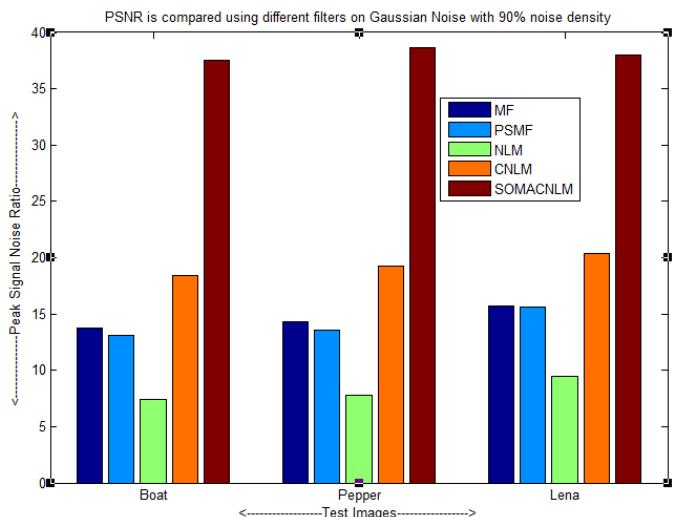


Fig. 14. Comparative chart of PSNR performance of different filters given in Table 3.

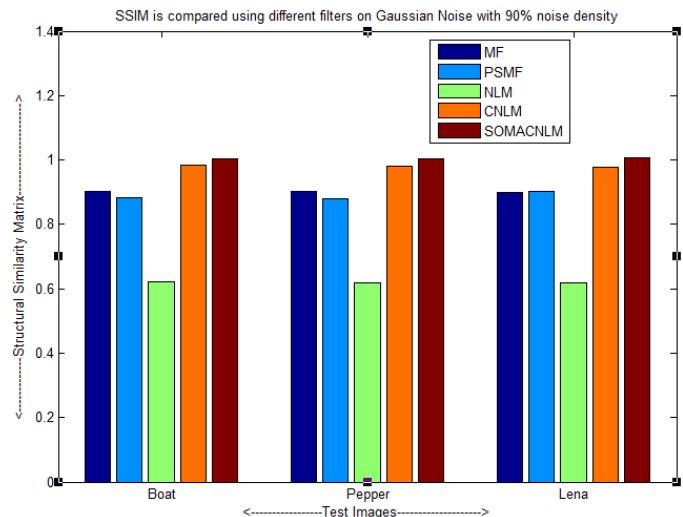


Fig. 15. Comparision chart of SSIM performance of dissimilar filters specified in Table 3.

IX. CONCLUSION

In this paper define an algorithm SOMACNLM for image denoising. The proposed technique finds the pixels similarity in the degradations imaginary depend on the numerous curvelet levels images produced and the degradations picture in keeping with the conjecture noise well-known deviation in noisy picture. The simulations have established that the proposed set of rules better the kingdom-of-art denoising processes in noise elimination phrases and element protection due to its efficiency in calculating pixel comparison and most suitable answer. This research proposed and discovers a new concept for image denoising exploiting SOMA and CNLM. There are two essential steps on this set of rules: first is Noise finding and second is Noise Removal. In noise detection step,

the idea of minimum and maximum of degradations pixels in a picture is used which offers higher noise finding capability and effectiveness. The experimental outcome presented which the proposed method of SOMACNLM is considerably superior various state-of-the-art schemes, both quantitatively and visually. This research proved that the image quality from 37.49% to 39.28% for 40% noise density. We have shown in this paper that SOMA is a good set of rules to locate a surest set of parameters for CNLM denoising. The simulation results, and associated evaluation criteria represent that our method generates good results, much better than existing work. It helps in noise removal along with preservation of fine details much better than that obtained with other methods.

This work can be further extended to optimization algorithm use. Optimization techniques may also be used to get better convergence problems and quality of solution. Also this work can be implementing on different type of noise and satellite images.

REFERENCES

- [1]. Nilima A. Bandane and DeekshaBhardwaj," Improved Hyperspectral Image Denoising Employing Sparse Representation", 2015 International Conference on Computational Intelligence and Communication Networks IEEE, pp: 475- 480.
- [2]. JingshaLv and Fuxiang Wang," Image Laplace Denoising based on Sparse Representation", 2015 International Conference on Computational Intelligence and Communication Networks, IEEE, pp: 373 377.
- [3]. Saroj K. Meher," Recursive and noise-exclusive fuzzy switching median filter for impulse noise reduction", Engineering Applications of Artificial Intelligence 30 (2014) 145–154
- [4]. Zhu Lin," A NONLOCAL MEANS BASED ADAPTIVE DENOISING FRAMEWORK FOR MIXED IMAGE NOISE REMOVAL", 2013 IEEE, pp: 454-458
- [5]. Xia Lan, Zhiyong Zuo, Random-valued impulse noise removal by the adaptiveswitching median detectors and detail-preserving regularization, Optik – Int. J.Light Electron. Opt. 125 (3) (2014) 1101–1105.
- [6]. Chandrika Saxena1 and Prof. Deepak Kourav," Noises and Image Denoising Techniques: A Brief Survey", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 3, March 2014, pp: 878- 885
- [7]. R.C. Gonzalez, R.E. Woods, Digital Image Processing, Prentice Hall, New Jersey,2002.
- [8]. Zhou Wang and David Zhang," Progressive Switching Median Filter for the Removal of Impulse Noise from Highly Corrupted Images", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: ANALOG AND DIGITAL SIGNAL PROCESSING, VOL. 46, NO. 1, JANUARY 1999, pp: 78-80
- [9]. I. Zelinka, SOMA-Self Organizing Migrating Algorithm, in G. Onwubolu, et al (Eds.), New optimization techniques in engineering, Springer, pp. 167-215, 2004
- [10]. Tomáš HORÁK and Pavel VAŘACHA," APPLICATION OF SELF-ORGANIZING MIGRATING ALGORITHM ON THE SHORTEST PATH PROBLEM", 7. - 9. 11. 2012, Jeseník, Czech Republic, EU
- [11]. P. Quan, Z. Lei, D. Guanzhong, and Z. Hongai, Two denoising methods by wavelet transform, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 47, issue: 12, pp. 3401-3406, 1999.
- [12]. Anupriya1 , Akash Tayal2 , "Wavelet based Image denoising using Self organizing migration algorithm",pp 1-6.
- [13]. Varsha.A and PreethaBasu," An Improved Dual Tree Complex Wavelet Transform based Image denoising using GCV thresholding", 2014 First International Conference on Computational Systems and

- Communications (ICCS) | 17-18 December 2014 | Trivandrum, pp:
133-138.
- [14]. Changying Dang, JianminGao, Zhao Wang Fumin Chen2 and Yulin Xiao," Optimized Wavelet Denoising Algorithm Using Hybrid Noise Model for Radiographic Images", 2014 Seventh International Symposium on Computational Intelligence and Design IEEE, pp: 144-149.
 - [15]. Ajay Kumar Boyat and Brijendra Kumar Joshi," Image Denoising using Wavelet Transform and Wiener Filter based on Log Energy Distribution over Poisson-Gaussian Noise Model", 2014 IEEE International Conference on Computational Intelligence and Computing Research.
 - [16]. K. Sasirekha and K. Thangavel," A Novel Wavelet based Thresholding for Denoising Fingerprint Image", 2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE) IEEE, pp: 119-124.
 - [17]. YihangLuoShengqian Wang, Chengzhi Deng, Jianping Xiao, Chao Long," Infrared Image Denoising Via L1/2 Sparse Representation Over Learned Dictionary", 2014 Seventh International Symposium on Computational Intelligence and Design IEEE, pp: 324-327.
 - [18]. FerielRomdhane, FaouziBenzarti and Hamid Amiri," 3D Medical Images Denoising", IEEE IPAS'14: INTERNATIONAL IMAGE PROCESSING APPLICATIONS AND SYSTEMS CONFERENCE 2014
 - [19]. MoussaOlfa and KhliifaNawres," Ultrasound Image Denoising using a Combination of Bilateral Filtering and Stationary Wavelet Transform", IEEE IPAS'14: INTERNATIONAL IMAGE PROCESSING APPLICATIONS AND SYSTEMS CONFERENCE 2014.
 - [20]. Sasaki S and SudhakarPutheti," Interpolation of CFA Color Images with Hybrid Image Denoising", 2014 Sixth International Conference on Computational Intelligence and Communication Networks IEEE,pp: 193-197.
 - [21]. XianjiuGuo and ChunyunGeng," Microalgae Image Denoising Algorithm Based on Stack", Proceeding of the 11th World Congress on Intelligent Control and Automation Shenyang, China, June 29 - July 4 2014, pp: 5088-5091.
 - [22]. N. Hemalatha," Image Denoising and Deblurring Using Non-Local Means Algorithm in Monochrome Images", International Journal of Engineering Research and General Science Volume 2, Issue 2, Feb-Mar 2014, pp: 197- 204.
 - [23]. Kaizhi Wu,a,b, Xuming Zhang,a, Mingyue Ding,a,, Curvelet based nonlocal means algorithm for image denoising", Int. J. Electron. Commun. (AEÜ) 68 (2014) 37– 43.

Discovery of Jumping Emerging Patterns Using Genetic Algorithm

Sumera Qurat ul Ain ^{#1}, Saif ur Rehman ^{*2}

UIIT, Arid Agriculture University,

Rawalpindi, Pakistan

[#]sumera_quratulain@yahoo.com

UIIT, Arid Agriculture University,

Rawalpindi, Pakistan

^{*}saif@uaar.edu.pk

Abstract— Patterns that only occur in objects belonging to a single class are called Jumping Emerging Patterns (JEP). JEP based Classifiers are considered one of the successful classification systems. Due to its comprehensibility, simplicity and strong differentiating abilities JEPs have captured significant recognition. However, discovery of JEPs in a large pattern space is normally a time consuming and challenging task because of their exponential behaviour. In this work a novel method based on genetic algorithm (GA) is proposed to discover JEPs in large pattern space. Since the complexity of GA is lower than other algorithms, so we have combined the power of JEPs and GA to find high quality JEPs from datasets to improve performance of classification system. Our proposed method explores a set of high quality JEPs from pattern search space unlike other methods in literature that compute complete set of JEPs. Large numbers of duplicate and redundant JEPs are filtered out during their discovery process. Experimental results show that our proposed Genetic-JEPs are effective and accurate for classification of a variety of data sets and in general achieve higher accuracy than other standard classifiers.

Keywords- Classification, Jumping Emerging Patterns, Genetic Algorithm.

I. INTRODUCTION

One of the primary objectives in data mining is classification and has been researched broadly in the fields of expert systems [1], machine learning [6], and neural networks [7] over decades. With reference to machine learning classification is the procedure to explore classifiers or predictive models that can differentiate testing data between different classes, in model derivation first training data is examined on the basis of quantifiable feature set (types include ordinal, categorical, integer-valued and real-valued) [8]. This training data contains information about the class of each instance from which it belongs. An essential part of the classification is termed as “classifiers” which is an implementation of an algorithm to classify unlabeled instances or a mathematical function which assigns test instances, to a category or class.

There exist problems when high accuracy is required along with explanation on which basis; it classifies objects to a class so that it should be understandable by users [2]. Many

classifiers lack such explanations, which is an important drawback to use them. For Example, in US Equal credit opportunity act, if a credit has been denied by a financial institution, an explanation also required explaining the reasons on which ground credit has been rejected to an applicant, if such reasons are not explained properly, credit denied considered illegal [4]. Medical diagnosis and mineral prospect ions are some other fields to name a few, where explanations and clarification are key user requirements in the classifiers [3]. A family in understandable classifiers are develop from emerging patterns [5]. “A pattern can be defined as an expression” in a language to describe a group of objects[7], moreover emerging pattern (EP) occurs abundantly in objects in one class except rare to locate in objects of other classes [9]. Jumping Emerging Pattern (JEP) is a type of EPs, are Itemset whose support rises suddenly from zero in one class of data to nonzero in another class, the ratio of support rises to infinite [1]. Due to their firm predictive power JEPs are extensively used in emerging pattern based classifiers [10]. Moreover JEPs are combination of simple conditions (e.g [Color = red] AND [Gender = female] AND [Age > 26]) therefore they are easily understandable by the user [11].

With all their merits of JEPs based classification, major drawback lies in the number of discovered patterns, useful for classification. If data has noise, number of patterns becomes larger. Searching useful JEPs in training data is a key procedure in JEP based classification system. Previous techniques either uses support threshold value or predefined numbers of patterns [12]. Due to these problems JEPs discovery is considered as a challenging task. To avoid redundancy in JEPs, we uses a subset of JEPs called minimal JEPs. Genetic algorithm adopts global search method that is better in finding solution, particularly in large search spaces. Since, time complexity of GA is less than other algorithms [13], so we have combined the power of JEPs and GA to find significant JEPs from different data sets for classification of data.

In our work we propose a new method to discover minimal JEPs that have significant affect on classification accuracy. Exploring fittest JEPs from datasets is a multistep process. In

first step, chromosomes(candidate solutions) are initiated these chromosomes are termed as population in literature, in subsequent steps solution of each population are used to generate new population in next stage having hope to get better population in each generation when compare with previous ones. This population generation is regulated with the help of a fitness function whole process is repeated until terminating conditions reached. After complete iterations of genetic algorithm, we will get a high quality fittest JEPs list, which will be evaluated on some test data for classification accuracy.

This paper is organized as: section II presents some background knowledge and related work. Section III presents proposed algorithm for discovery of JEPs. Section IV presents experimental results and analysis and section V describes conclusion.

II. BACKGROUND AND RELATED WORK

This section consists of some basic concepts and definitions, used throughout in this paper. A dataset is a set of objects having multiple attributes (A_1, A_2, \dots, A_n), each data object is referred as an instance. Each instance of dataset has associated class $C \in \{C_1, C_2, \dots, C_n\}$ [14]. If I represents set of items or patterns, P is an itemset which is subset of I , if any instance S holds an itemset P provided $P \subseteq S$ then support of an itemset P in a data set D represent as $supp_D(P)$ is $count_D(P)/|D|$, Where $Count_D(P)$ is the number of instances in D containing P , and $|D|$ denoted total instances in D . Growth Rate is the ratio of support of a pattern in its native class C_p with the support of pattern in other class, it represents the predictive power of the pattern. $Growth\ Rate(P)$ is defined as:

$$Growth\ Rate(P) = \begin{cases} 0, & \text{if } support(P, C) = 0 \wedge support(P, C_p) = 0 \\ \infty, & \text{if } support(P, C) = 0 \wedge support(P, C_p) > 0 \\ \max\left(\frac{support(P, C_p)}{support(P, C)}, \frac{support(P, C)}{support(P, C_p)}\right), & \text{otherwise} \end{cases}$$

Definition 1. Let D is a dataset comprises on D_1 and D_2 where objects of D_1 belongs to one class and objects of D_2 belongs to a different class, a Jumping Emerging Pattern (JEP) from D_1 to D_2 is an itemset X , that satisfies the condition:

$$supp_{D_1}(X)=0 \text{ and } supp_{D_2}(X) > 0$$

Definition 2. If a JEP does not hold another JEP as a proper subset, then it is called minimal JEP.

Example 1. Let For dataset presented in Table I, itemsets $\{l,p\}(1,0)$, $\{l,n,p\}(1,0)$, $\{l,o,p\}(1,0)$, $\{l, n, o, p\}(1,0)$, $\{m, p\}(2,0)$, $\{m, n, p\}(1,0)$, $\{m, o, p\}(1,0)$, $\{m, n, o, p\}(1,0)$, and $\{n, o, p\}(2,0)$ are JEPs of class 1; itemsets $\{l, m\}(0,2)$, $\{l, m, n\}(0,1)$, $\{l, m, o\}(0,1)$, and $\{l, m, n, o\}(0,1)$ are JEPs of class 2. The total number of JEPs is 13, and four among them are minimal JEPs, namely, $\{l, p\}(1 : 0)$, $\{m, p\}(2 : 0)$, $\{n, o, p\}(2 : 0)$, and $\{l, m\}(0 : 2)$ [1].

In available literature, researchers proposed several JEP-based Classifiers [15]. In [16] concept of Essential Jumping Emerging pattern (eJEPs) is proposed, this classifier utilizes

less JEPs than the other JEP classifier. Limitations of this research includes that it is not complete not sound to discover all JEPs and yields itemsets other than actual JEPs [2]. Classifier proposed in [2] is a FP tree based technique which

TABLE I
Example dataset having two classes

ID	Class Labels	Instances (Itemsets)
1	D_1	$\{l,n,o,p\}$
2	D_1	$\{l\}$
3	D_1	$\{m,p\}$
4	D_1	$\{m,n,o,p\}$
5	D_2	$\{l,m\}$
6	D_2	$\{n,p\}$
7	D_2	$\{l,m,n,o\}$
8	D_2	$\{o,p\}$

uses a FP growth algorithm to explore JEPs, exploring tree based structure has high time and space complexity when comparing with other data structures, in [1] Strong JEP is proposed, authors claim that SJEPs achieve higher quality when comparing with JEPs, it is an efficient classifier due to the less number of SJEPs. In same research author proposed Noise tolerant emerging patterns (NEPs) and generalized noise tolerant emerging patterns (GNEPs).The limitations of these JEPs include that they find SJEPs in two steps initially it generate a large number of JEPs then in second step generated JEPs are filtered out to produce a small set of SJEPs. This twostep process is time consuming and its time and space complexity is higher than its competitors in literature, another major demerit of this study is its requirement of minimum threshold value to prune large set of JEPs. In [17] negJEP-Classifier and JEPN-classifier introduced, these classifiers use negative information for classification, results revealed that a limitation of this research is its infeasibility to build JEPN classifier for some data sets [17]. [18] discover Top K minimal jumping emerging patterns, proposed method finds strong JEPs which are CP tree based, instead of exploring all JEPs it reduces search space and explore top K - JEPs only on the bases of increasing minimum threshold support value. The drawback of this research includes additional pattern counting [18]. In [19] Highest impact JEPs introduced, which is based on introducing a new coefficient REAL/ALL, this coefficient is helpful in comparing discriminative power of distinguished JEPs collections, demerit of this research is its limited application to specific datasets. An improved tree based method is proposed in [14], this method explores SJEPs more efficiently, drawback of this method include it cannot mine SJEPs directly, [20] extended border based mining method and proposed a classifying model for discovering JEPs with occurrence count for classification. The drawback of this method which can be improved is its discovery of a large number of JEPs. The scheme presented in [21] is based on CP tree whose growth rate is dynamically increasing, it prunes the search space with the help of new pattern pruning method, demerits of this research is like in every tree based search is its space complexity and inability to handle more complex patterns [21]. In [12] a method for discovery of minimal JEPs introduced and to the best of my knowledge I found the last

reported work on JEPs is the use of cosine similarity with JEPs in classification by [10] but its accuracy is compromised.

Various methods have been proposed in the past to reduce the huge pattern set but it is still an open research area to considerably reduce the number of derived patterns and improve the worth of selected patterns.

III. PROPOSED ALGORITHM FOR DISCOVERY OF JEPs

Inspired by Darwin's Evolution theory Genetic algorithm is a solution to many problems. Genetic algorithm adopts global search strategy that is better for searching optimized solution, particularly in bigger search spaces. In the following section, detailed procedure of discovery of JEPs by using Genetic Algorithm is given, major steps includes: Individual representation, Initial Population Generation, Fitness evaluation, creation of next generation by using selection, reproduction, crossover and mutation.

A. Individual Representation

Different type of attributes exists, e.g. categorical, numerical, in numerical attributes an assumption is made, this assumption is about the discretized range of value in intervals. Different combinations of attribute and their values are normally used to express patterns, like (*Color = red, gender = female, Age = 26*) or as logical properties, like [*Color = red*] AND [*Gender = female*] AND [*Age > 26*]. Normally, data need to be encoded into chromosomes because genetic algorithm cannot directly handle data in the solution search space. Every Chromosome represents a candidate pattern and have the form of ($A_1 = V_{1j} \wedge \dots \wedge A_n = V_{nj}$). Where A_i represents i^{th} feature and V_{ij} represents j^{th} value of the i^{th} feature's domain and n represents length of chromosome [22].

A_1	A_2	A_3	A_4	A_5	A_n
V_{1j}	V_{2j}	V_{3j}	V_{4j}	V_{5j}	V_{nj}

Fig. 1. Individual Representation

B. Initial Population Generation

We have randomly created population of n chromosomes (where n is the number of chromosomes in population), population size can be fixed according the problem. In our method we have used population of 100 chromosomes. These chromosomes represent patterns by randomly selecting attributes and their values.

C. Fitness Evaluation of Individual

When applying Genetic Algorithm fitness function is the central part that plays a vital role in problem optimization. We need to find some measures (*support* and *growth ratio* described in section II) of these patterns to analyze their differentiating power between different classes. After calculating the *support* and *growth ratio* values of a pattern proposed algorithm will check the pattern whether it is a JEP or not? In our method we are exploring Minimal JEPs so extra attributes will be removed by applying a pruning method. Growth ratio of a pattern will be its Fitness value. Having very low fitness value patterns will be less probable to survive for

the next generation. Class will be assigned to the pattern according to its growth rate. Then discovered Minimal JEPs will be added to a global pattern list. Each time when we add a discovered pattern to the pattern list there is a check to avoid duplication of pattern in global pattern list.

D. Create Next generation

After fitness evaluation we will create new population for the next generation by repeating following steps unless new population according to population size is completed.

1) *Selection*: In selection process those patterns are used which has good fitness values to produce next generation with the expectation to have offspring that has higher fitness value than their parents. In order to getting better members we have used Proportional Roulette wheel selection. Each chromosome has selection probability i.e. directly proportional with individual's fitness value. Chromosomes with higher probability capture the larger fragment; while the less fit captures likewise smaller fragment in the roulette wheel. Clearly those chromosomes which have larger fragment size have more chances to be chosen as parent for next generation. Let f_1, f_2, \dots, f_n be fitness values of individual 1, 2, ..., n. Then the selection probability, P_i for individual i can be calculated as[23]:

$$P_i = \frac{f_i}{\sum_{j=1}^n f_j}$$

2) *Crossover*: Procedure of crossover begins with selecting more than two parent solutions (patterns) and generating two child solutions from them. We have used single point crossover so first we select a random point then beyond that point all data in both patterns is swapped between the two parents. This results in two offspring patterns.

3) *Mutation*: Mutation will be applied on a single chromosome at a time. It preserves gene diversity in population and ensures searching in the entire solution space. In our proposed method mutation rate is very low i.e 5 % because we don't want major changes in population it is used only for global optimization. Complete algorithm is given as follows:

Algorithm 1: Algorithm for discovery of JEPs using GA

Input: D : m x n matrix, training dataset.

GenNo : maximum number of allowed generations

MR : Mutation rate

CR : Crossover Rate

PSize : Maximum number of chromosomes in

Population

Output: The best Jumping emerging pattern in all generations.

1. Generate random population of n chromosomes , where $n = 1 \dots PSize$.
2. **for** $i = 1$ to GenNo **do**
3. **for** $j = 1$ to PSize **do**
4. Calculate fitness of each individual.
5. Add discovered patterns to list.
6. Prune pattern list.

7. Calculate selection probability of each individual.
8. **end do**
9. Create fresh population through repeating the following steps unless the population completion:
 Selection().
 Crossover();
 Mutation();
10. Add new child to fresh population.
11. Use newly created population for further running of algorithm.
12. **Check Termination criteria:** If the terminating criterion is satisfied, **stop**. Termination condition is predefined number of generations. If termination criterion is not satisfied then go to **step 2**.
13. **endfor**
14. Evaluate Pattern list on test set.
15. Display Result.
16. end

Classification may be binary or multiclass. In binary classification, test objects can be classified in two classes only, whereas in multi-class more than two classes can be assigned to test objects. Our proposed method can be used for multiclass classification. For multiclass classification we will use one against all technique.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Databases used in the experiments are downloaded from the UCI ML repository. Description of Datasets is given in the table II. Results of all the experimental are expressed as Average, i.e. ratio of correct classified instances with the overall number of instances present in database. To ensure accuracy CV-10 methodology is used. Results represent the average classification performance over the 10 folds. Moreover, frequency of experiment execution is five. To ensure robustness in the values, results of all five executions are than used to calculate average.

TABLE II
Description of Datasets

Datasets	# of Instances	# of attributes	# of classes
adult+stretch_data	20	5	2
breast-cancer	286	10	2
Crx	690	16	2
Diabete	768	9	2
Glass	214	10	6
Hayes Roth	132	5	3
Iris	150	5	3
Monks	432	7	2
Nursery	12960	9	5
Sonar	208	61	2
Soybean-small	47	22	4
Tic Tac Toe	958	10	2
Vehicle	846	19	4
Xad	94	19	4
Xae	94	19	4

It is evident that all these experiments are aimed to assess the success and predictive power of the discovered JEPs and its discovery method. For the experiments of the classifier, 15 common and benchmark databases for classification are used. Proposed technique is implemented in Microsoft Visual C# .Net Framework 4.0, using Microsoft Visual Studio 2015. All experiments were conducted on HP ProBook 4530 Core i5-2430M CPU @ 2.40 GHz, 4Gb RAM) running Windows 7 Home Premium 64 bit version. Statistical details on important information of the data set used for the experiments are given in the following table.

In experiments, different parameters have been used for the evaluation purposes. These parameters are included *GenNo* represents the Total Number of Generation used for the iteration of the proposed algorithm. Other parameters are included *MR* which is the mutation rate for global optimizaiotn. *CR* is the Cross Over Rate which is used for the generation of the next population. *PSize* is the size of population. Finally, the *D* parameter is used for training dataset given to proposed algorithm. We have used different values for the above mentioned parameters. These parameter values are given in the Table III.

TABLE III
Experimental Parameters

Parameter	Value
GenNo	50
MR	0.15
CR	0.85
PSize	100
D	M X N matrix training data set

Genetic-JEPs (G-JEPs) accuracy has been compared with the seven popular classifiers, i.e. Naïve Bayes, J48, ID3, Zero-R, One- R, bagging, and Random Forest. The accuracy of Naive Bayes, J48, ID3, Zero-R, One- R, bagging, and Random Forest are obtained using Weka 3.6.9 implementation.

TABLE IV
G-JEPs Accuracy on various datasets

Datasets	G-JEPs Accuracy
adult+stretch_data	100.00
breast-cancer	96.494
Crx	85.285
Diabete	88.34
Glass	59.345
Hayes-roth	100.00
Iris	95.066
Monks	99.459
Nursery	90.848
Sonar	57.211
Soybean-small	100.00
tic-tac-toe	98.121
vehicle	56.146
xad	71.276
xae	67.021

We present a best/ same / poor summary in table VI to evaluate performance of the G- JEPs against competitors. best/same/poor represents the number of data sets on which

the G-JEPs results greater accuracy than others, the number for which the both classifier results same accuracy, and the number on which comparator achieves greater accuracy. From table V and table VI, it is clearly shown that the G-JEPs obtained significantly better accuracy on majority datasets.

TABLE V
Accuracy comparison with standard classifiers

Datasets	Naive Bayes	J48	bagging	ID3	Random Forest	ZeroR	One-R	G-JEPs
adult+stretch_data	100.00	100.00	100.00	100.00	100.00	60.00	70.00	100.00
breast-cancer	72.08	75.524	67.832	56.993	69.231	70.279	65.734	96.494
crx	84.787	85.36	85.942	72.029	83.913	55.507	85.507	85.285
diabete	72.91	74.218	73.046	63.411	69.531	65.104	74.088	88.34
glass	63.081	59.813	63.084	61.215	69.53	35.514	44.392	59.345
Hayes-roth	80.303	72.727	77.272	62.121	77.272	37.878	43.930	100.00
iris	87.333	90.667	90.00	89.333	90.00	33.333	82.667	95.066
monks	75.00	96.527	98.6111	95.370	96.527	49.074	75.00	99.459
nursery	90.324	97.052	97.268	98.186	97.938	33.333	70.972	90.848
sonar	67.788	71.153	77.884	---	79.326	53.365	62.50	57.211
Soybean-small	100.00	97.872	100.00	95.744	100.00	36.170	87.234	100.00
tic-tac-toe	69.624	85.073	90.710	83.298	90.709	65.344	69.937	98.121
vehicle	57.10	65.484	67.021	61.820	65.484	25.650	41.489	56.146
xad	56.383	53.191	48.936	38.297	57.446	30.851	41.489	71.276
xae	57.447	53.191	48.936	38.297	57.446	30.851	41.489	67.021

TABLE VI
Best/ same/ poor record of G-JEPs vs. Alternatives

JEPs Vs	Naive Bayes	J48	bagging	ID3	Random Forest	0- R	1- R
best/same/poor	10,2,3	9,1,5	8,2,5	11,1,3	9,2,4	15,0,0	13,0,2

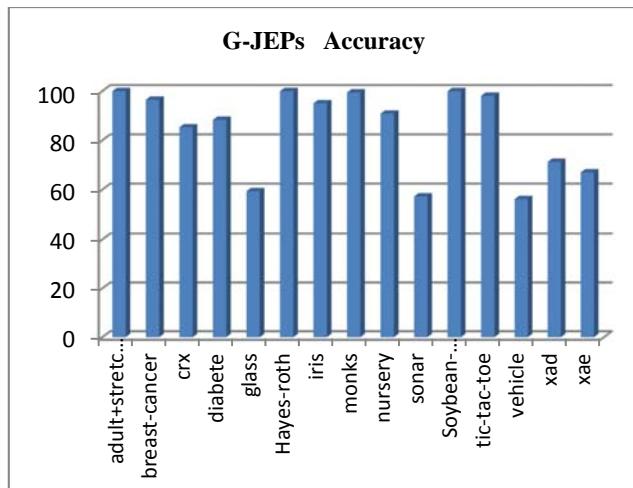


Fig. 2. G-JEPs accuracy on various UCI datasets

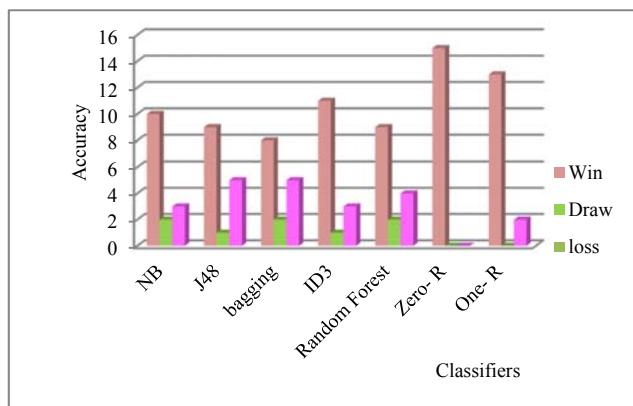


Fig. 3. Best/ Same/ Poor performance of G-JEPs vs Alternatives

V. CONCLUSION

A comprehensive study about Jumping Emerging Patterns (JEPs) and the earlier related research works has been presented to highlight their limitations in order to address classification problems. Our contribution presented in this work is a novel method for JEPs discovery using Genetic Algorithm aimed to overcome the deficiencies of the previous related approaches in this domain. To decrease the number of JEPs, we used a function for pruning. Large numbers of duplicate and redundant JEPs are filtered out during their discovery process. We developed an accurate and effective classification model based on JEPs. We have performed experiments on standard data sets to illustrate that proposed G-JEPs are effective and accurate for classification of a variety of data sets and in general achieves higher accuracy than other standard classifiers.

REFERENCES

- [1] H. Fan, and k. Ramamohanarao. 2006. Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers, *knowledge and Data Engineering, IEEE Transactions on*, vol. 18, pp.721-737 2006.
- [2] J. Bailey, T. Manoukian, and K. Ramamohanarao, "Fast Algorithms for Mining Emerging Patterns", *Proc. Sixth European Conf. Principles and Practice of Knowledge Discovery in Databases(PKDD'02)*, 2002.
- [3] M. Garcia-Borroto, J. F. Martínez-Trinidad and J. A. Carrasco-Ochoa, 2014. A survey of emerging patterns for supervised classification, *Artificial Intelligence Review*, 42(4), pp. 705-721.
- [4] M. García-Borroto, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, M.A. Medina-Pérez, and J. Ruiz-Shulcloper. 2010. LCMine: An efficient algorithm for mining discriminative regularities and its application in supervised classification. *Pattern Recognition*, 43(9), pp. 3025-3034.
- [5] K. Ramamohanarao and H. Fan. 2007. Patterns based classifiers, *World Wide Web*, 10(1), pp.71-83.
- [6] T.M. Mitchell, Machine Learning. McGraw-Hill Higher Education, 1997.
- [7] G. Pateski, and W.Frawley. 1991. Knowledge discovery in databases, *MIT press*.
- [8] Y. Ma, L. Bing and H.Wynne. 1998. Integrating classification and association rule mining, In *Proceeding of the fourth international conference on knowledge discovery and data mining*.
- [9] P. Andruszkiewicz. 2011. Lazy approach to preserving classification with emerging patterns, In *Emerging intelligent technologies in industry*, pp. 253-268. Springer Berlin Heidelberg.
- [10] M. Ferrandin, A. Boava, and A.S. R. Pinto. 2015. Classification Using Jumping Emerging Patterns and Cosine Similarity. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pp. 682.
- [11] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In: KDD, pp.43-52n(1999).
- [12] B. Kane, B. Cuissart and B. Cremilleux. 2015. Minimal Jumping Emerging Patterns: Computation and Practical Assessment, In *Advances in Knowledge Discovery and Data Mining*, pp. 722-733. Springer International Publishing.
- [13] M. Kabir, M. J., S. Xu, B. H. Kang, and Z. Zhao. 2015. Comparative analysis of genetic based approach and Apriori algorithm for mining maximal frequent item sets. In *IEEE Congress on Evolutionary Computation (CEC)*, pp. 39-45.
- [14] Chen, Xiangtao, and Lijuan Lu. "An improved algorithm of mining Strong Jumping Emerging Patterns based on sorted SJEP-Tree." *Bio-Inspired Computing: Theories and Applications (BIC-TA)*, 2010 IEEE Fifth International Conference on. IEEE, 2010.
- [15] Li, J., G. Dong and K. Ramamohanarao. 2001. Making use of the most expressive jumping emerging patterns for classification, *Knowledge and Information systems*, 3(2), pp. 131-145.
- [16] H. Fan and K. Ramamohanarao. 2002. An efficient single-scan algorithm for mining essential jumping emerging patterns for classification, In *Advances in Knowledge Discovery and Data Mining*, Springer Berlin Heidelberg, pp. 456-462.
- [17] P. Terlecki and K. Walczak. 2007. Jumping emerging patterns with negation in transaction databases-Classification and discovery, *Information Sciences*, 177(24), pp. 5675-5690.
- [18] P. Terlecki and K. Walczak. 2008. Efficient discovery of top-k minimal jumping emerging patterns, In *Rough Sets and Current Trends in Computing*, Springer Berlin Heidelberg, pp. 438-447.
- [19] T. Gambin, and K. Walczak. 2009. Classification based on the highest impact jumping emerging patterns, In *Computer Science and Information Technology, IMCSIT'09, International Multiconference on*, pp. 37-42. IEEE.
- [20] Walczak. 2009. Classification based on the highest impact jumping emerging patterns, In *Computer Science and Information Technology, IMCSIT'09, International Multiconference on*, pp. 37-42. IEEE.
- [21] Kobyliński, L. and K. Walczak. 2011. Efficient mining of jumping emerging patterns with occurrence counts for classification, *In Transactions on rough sets XIII*, Springer Berlin Heidelberg, pp. 73-88.
- [22] Liu, Q., P. Shi, Z. Hu, and Y. Zhang. 2014. A novel approach of mining strong jumping emerging patterns based on BSC-tree. *International Journal of Systems Science*, 45(3), pp. 598-615.
- [23] X. Shi, J. and H. Lei. 2008. A genetic algorithm-based approach for classification rule discovery. International Conference on Information Management, Innovation Management and Industrial Engineering, Vol. 1, pp. 175-178. IEEE..

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore, Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

Prof. Syed S. Rizvi, University of Bridgeport, USA
Dr. Shahbaz Pervez Chattha, University of Engineering and Technology Taxila, Pakistan
Dr. Shishir Kumar, Jaypee University of Information Technology, Wakanaghata (HP), India
Dr. Shahid Mumtaz, Portugal Telecommunication, Instituto de Telecomunicações (IT) , Aveiro, Portugal
Dr. Rajesh K Shukla, Corporate Institute of Science & Technology Bhopal M P
Dr. Poonam Garg, Institute of Management Technology, India
Dr. S. Mehta, Inha University, Korea
Dr. Dilip Kumar S.M, Bangalore University, Bangalore
Prof. Malik Sikander Hayat Khiyal, Fatima Jinnah Women University, Rawalpindi, Pakistan
Dr. Virendra Gomase , Department of Bioinformatics, Padmashree Dr. D.Y. Patil University
Dr. Irraivan Elamvazuthi, University Technology PETRONAS, Malaysia
Dr. Saqib Saeed, University of Siegen, Germany
Dr. Pavan Kumar Gorakavi, IPMA-USA [YC]
Dr. Ahmed Nabih Zaki Rashed, Menoufia University, Egypt
Prof. Shishir K. Shandilya, Rukmani Devi Institute of Science & Technology, India
Dr. J. Komala Lakshmi, SNR Sons College, Computer Science, India
Dr. Muhammad Sohail, KUST, Pakistan
Dr. Manjaiah D.H, Mangalore University, India
Dr. S Santhosh Baboo, D.G.Vaishnav College, Chennai, India
Prof. Dr. Mokhtar Beldjehem, Sainte-Anne University, Halifax, NS, Canada
Dr. Deepak Laxmi Narasimha, University of Malaya, Malaysia
Prof. Dr. Arunkumar Thangavelu, Vellore Institute Of Technology, India
Dr. M. Azath, Anna University, India
Dr. Md. Rabiul Islam, Rajshahi University of Engineering & Technology (RUET), Bangladesh
Dr. Aos Alaa Zaidan Ansaef, Multimedia University, Malaysia
Dr Suresh Jain, Devi Ahilya University, Indore (MP) India,
Dr. Mohammed M. Kadhum, Universiti Utara Malaysia
Dr. Hanumanthappa. J. University of Mysore, India
Dr. Syed Ishtiaque Ahmed, Bangladesh University of Engineering and Technology (BUET)
Dr Akinola Solomon Olalekan, University of Ibadan, Ibadan, Nigeria
Dr. Santosh K. Pandey, The Institute of Chartered Accountants of India
Dr. P. Vasant, Power Control Optimization, Malaysia
Dr. Petr Ivankov, Automatika - S, Russian Federation
Dr. Utkarsh Seetha, Data Infosys Limited, India
Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal
Dr. (Mrs) Padmavathi Ganapathi, Avinashilingam University for Women, Coimbatore
Assist. Prof. A. Neela madheswari, Anna university, India
Prof. Ganesan Ramachandra Rao, PSG College of Arts and Science, India
Mr. Kamanashis Biswas, Daffodil International University, Bangladesh
Dr. Atul Gonsai, Saurashtra University, Gujarat, India
Mr. Angkoon Phinyomark, Prince of Songkla University, Thailand
Mrs. G. Nalini Priya, Anna University, Chennai
Dr. P. Subashini, Avinashilingam University for Women, India
Assoc. Prof. Vijay Kumar Chakka, Dhirubhai Ambani IICT, Gandhinagar ,Gujarat
Mr Jitendra Agrawal, : Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal
Mr. Vishal Goyal, Department of Computer Science, Punjabi University, India
Dr. R. Baskaran, Department of Computer Science and Engineering, Anna University, Chennai

Assist. Prof. Kanwalvir Singh Dhindsa, B.B.S.B.Engg. College, Fatehgarh Sahib (Punjab), India
Dr. Jamal Ahmad Dargham, School of Engineering and Information Technology, Universiti Malaysia Sabah
Mr. Nitin Bhatia, DAV College, India
Dr. Dhavachelvan Ponnurangam, Pondicherry Central University, India
Dr. Mohd Faizal Abdollah, University of Technical Malaysia, Malaysia
Assist. Prof. Sonal Chawla, Panjab University, India
Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia
Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia
Professor Dr. Sabu M. Thampi, .B.S Institute of Technology for Women, Kerala University, India
Mr. Noor Muhammed Nayeem, Université Lumière Lyon 2, 69007 Lyon, France
Dr. Himanshu Aggarwal, Department of Computer Engineering, Punjabi University, India
Prof R. Naidoo, Dept of Mathematics/Center for Advanced Computer Modelling, Durban University of Technology, Durban, South Africa
Prof. Mydhili K Nair, Visweswaraiah Technological University, Bangalore, India
M. Prabu, Adhiyamaan College of Engineering/Anna University, India
Mr. Swakkhar Shatabda, United International University, Bangladesh
Dr. Abdur Rashid Khan, ICIT, Gomal University, Dera Ismail Khan, Pakistan
Mr. H. Abdul Shabeer, I-Nautix Technologies, Chennai, India
Dr. M. Aramudhan, Perunthalaivar Kamarajar Institute of Engineering and Technology, India
Dr. M. P. Thapliyal, Department of Computer Science, HNB Garhwal University (Central University), India
Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
Mr. Zeashan Hameed Khan, Université de Grenoble, France
Prof. Anil K Ahlawat, Ajay Kumar Garg Engineering College, Ghaziabad, UP Technical University, Lucknow
Mr. Longe Olumide Babatope, University Of Ibadan, Nigeria
Associate Prof. Raman Maini, University College of Engineering, Punjabi University, India
Dr. Maslin Masrom, University Technology Malaysia, Malaysia
Sudipta Chattopadhyay, Jadavpur University, Kolkata, India
Dr. Dang Tuan NGUYEN, University of Information Technology, Vietnam National University - Ho Chi Minh City
Dr. Mary Lourde R., BITS-PILANI Dubai, UAE
Dr. Abdul Aziz, University of Central Punjab, Pakistan
Mr. Karan Singh, Gautam Budtha University, India
Mr. Avinash Pokhriyal, Uttar Pradesh Technical University, Lucknow, India
Associate Prof Dr Zuraini Ismail, University Technology Malaysia, Malaysia
Assistant Prof. Yasser M. Alginahi, Taibah University, Madinah Munawwarah, KSA
Mr. Dakshina Ranjan Kisku, West Bengal University of Technology, India
Mr. Raman Kumar, Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India
Associate Prof. Samir B. Patel, Institute of Technology, Nirma University, India
Dr. M. Munir Ahmed Rabbani, B. S. Abdur Rahman University, India
Asst. Prof. Koushik Majumder, West Bengal University of Technology, India
Dr. Alex Pappachen James, Queensland Micro-nanotechnology center, Griffith University, Australia
Assistant Prof. S. Hariharan, B.S. Abdur Rahman University, India
Asst Prof. Jasmine. K. S, R.V. College of Engineering, India
Mr Naushad Ali Mamode Khan, Ministry of Education and Human Resources, Mauritius
Prof. Mahesh Goyani, G H Patel College of Engg. & Tech, V.V.N. Anand, Gujarat, India
Dr. Mana Mohammed, University of Tlemcen, Algeria
Prof. Jatinder Singh, Universal Institution of Engg. & Tech. CHD, India

Mrs. M. Anandhavalli Gauthaman, Sikkim Manipal Institute of Technology, Majitar, East Sikkim
Dr. Bin Guo, Institute Telecom SudParis, France
Mrs. Maleika Mehr Nigar Mohamed Heenaye-Mamode Khan, University of Mauritius
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Mr. V. Bala Dhandayuthapani, Mekelle University, Ethiopia
Dr. Irfan Syamsuddin, State Polytechnic of Ujung Pandang, Indonesia
Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
Mr. Ravi Chandiran, Zagro Singapore Pte Ltd. Singapore
Mr. Milindkumar V. Sarode, Jawaharlal Darda Institute of Engineering and Technology, India
Dr. Shamimul Qamar, KSJ Institute of Engineering & Technology, India
Dr. C. Arun, Anna University, India
Assist. Prof. M.N.Birje, Basaveshwar Engineering College, India
Prof. Hamid Reza Naji, Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran
Assist. Prof. Debasis Giri, Department of Computer Science and Engineering, Haldia Institute of Technology
Subhabrata Barman, Haldia Institute of Technology, West Bengal
Mr. M. I. Lali, COMSATS Institute of Information Technology, Islamabad, Pakistan
Dr. Feroz Khan, Central Institute of Medicinal and Aromatic Plants, Lucknow, India
Mr. R. Nagendran, Institute of Technology, Coimbatore, Tamilnadu, India
Mr. Amnach Khawne, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok, Thailand
Dr. P. Chakrabarti, Sir Padampat Singhania University, Udaipur, India
Mr. Nafiz Imtiaz Bin Hamid, Islamic University of Technology (IUT), Bangladesh.
Shahab-A. Shamshirband, Islamic Azad University, Chalous, Iran
Prof. B. Priestly Shan, Anna Univeristy, Tamilnadu, India
Venkatramreddy Velma, Dept. of Bioinformatics, University of Mississippi Medical Center, Jackson MS USA
Akshi Kumar, Dept. of Computer Engineering, Delhi Technological University, India
Dr. Umesh Kumar Singh, Vikram University, Ujjain, India
Mr. Serguei A. Mokhov, Concordia University, Canada
Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia
Dr. Awadhesh Kumar Sharma, Madan Mohan Malviya Engineering College, India
Mr. Syed R. Rizvi, Analytical Services & Materials, Inc., USA
Dr. S. Karthik, SNS Collegeof Technology, India
Mr. Syed Qasim Bukhari, CIMET (Universidad de Granada), Spain
Mr. A.D.Potgantwar, Pune University, India
Dr. Himanshu Aggarwal, Punjabi University, India
Mr. Rajesh Ramachandran, Naipunya Institute of Management and Information Technology, India
Dr. K.L. Shunmuganathan, R.M.K Engg College , Kavaraipettai ,Chennai
Dr. Prasant Kumar Pattnaik, KIST, India.
Dr. Ch. Aswani Kumar, VIT University, India
Mr. Ijaz Ali Shoukat, King Saud University, Riyadh KSA
Mr. Arun Kumar, Sir Padam Pat Singhania University, Udaipur, Rajasthan
Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia
Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA
Mr. Mohd Zaki Bin Mas'ud, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
Prof. Dr. R. Geetharamani, Dept. of Computer Science and Eng., Rajalakshmi Engineering College, India
Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
Dr. S. Abdul Khader Jilani, University of Tabuk, Tabuk, Saudi Arabia
Mr. Syed Jamal Haider Zaidi, Bahria University, Pakistan

Dr. N. Devarajan, Government College of Technology, Coimbatore, Tamilnadu, INDIA
Mr. R. Jagadeesh Kannan, RMK Engineering College, India
Mr. Deo Prakash, Shri Mata Vaishno Devi University, India
Mr. Mohammad Abu Naser, Dept. of EEE, IUT, Gazipur, Bangladesh
Assist. Prof. Prasun Ghosal, Bengal Engineering and Science University, India
Mr. Md. Golam Kaosar, School of Engineering and Science, Victoria University, Melbourne City, Australia
Mr. R. Mohammad Shafi, Madanapalle Institute of Technology & Science, India
Dr. F.Sagayaraj Francis, Pondicherry Engineering College, India
Dr. Ajay Goel, HIET, Kaithal, India
Mr. Nayak Sunil Kashibarao, Bahirji Smarak Mahavidyalaya, India
Mr. Suhas J Manangi, Microsoft India
Dr. Kalyankar N. V., Yeshwant Mahavidyalaya, Nanded, India
Dr. K.D. Verma, S.V. College of Post graduate studies & Research, India
Dr. Amjad Rehman, University Technology Malaysia, Malaysia
Mr. Rachit Garg, L K College, Jalandhar, Punjab
Mr. J. William, M.A.M college of Engineering, Trichy, Tamilnadu, India
Prof. Jue-Sam Chou, Nanhua University, College of Science and Technology, Taiwan
Dr. Thorat S.B., Institute of Technology and Management, India
Mr. Ajay Prasad, Sir Padampat Singhania University, Udaipur, India
Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology & Science, India
Mr. Syed Raifi Hussain, Ahsanullah University of Science and Technology, Bangladesh
Mrs Fazeela Tunnis, Najran University, Kingdom of Saudi Arabia
Mrs Kavita Taneja, Maharishi Markandeshwar University, Haryana, India
Mr. Maniyar Shiraz Ahmed, Najran University, Najran, KSA
Mr. Anand Kumar, AMC Engineering College, Bangalore
Dr. Rakesh Chandra Gangwar, Beant College of Engg. & Tech., Gurdaspur (Punjab) India
Dr. V V Rama Prasad, Sree Vidyanikethan Engineering College, India
Assist. Prof. Neetesh Kumar Gupta, Technocrats Institute of Technology, Bhopal (M.P.), India
Mr. Ashish Seth, Uttar Pradesh Technical University, Lucknow, UP India
Dr. V V S S Balaram, Sreenidhi Institute of Science and Technology, India
Mr Rahul Bhatia, Lingaya's Institute of Management and Technology, India
Prof. Niranjana Reddy. P, KITS, Warangal, India
Prof. Rakesh. Lingappa, Vijetha Institute of Technology, Bangalore, India
Dr. Mohammed Ali Hussain, Nimra College of Engineering & Technology, Vijayawada, A.P., India
Dr. A.Srinivasan, MNM Jain Engineering College, Rajiv Gandhi Salai, Thorapakkam, Chennai
Mr. Rakesh Kumar, M.M. University, Mullana, Ambala, India
Dr. Lena Khaled, Zarqa Private University, Aman, Jordan
Ms. Supriya Kapoor, Patni/Lingaya's Institute of Management and Tech., India
Dr. Tossapon Boongoen, Aberystwyth University, UK
Dr. Bilal Alatas, Firat University, Turkey
Assist. Prof. Jyoti Praaksh Singh, Academy of Technology, India
Dr. Ritu Soni, GNG College, India
Dr. Mahendra Kumar, Sagar Institute of Research & Technology, Bhopal, India.
Dr. Binod Kumar, Lakshmi Narayan College of Tech.(LNCT)Bhopal India
Dr. Muzhir Shaban Al-Ani, Amman Arab University Amman – Jordan
Dr. T.C. Manjunath, ATRIA Institute of Tech, India
Mr. Muhammad Zakarya, COMSATS Institute of Information Technology (CIIT), Pakistan

Assist. Prof. Harmunish Taneja, M. M. University, India
Dr. Chitra Dhawale , SICSR, Model Colony, Pune, India
Mrs Sankari Muthukaruppan, Nehru Institute of Engineering and Technology, Anna University, India
Mr. Aaqif Afzaal Abbasi, National University Of Sciences And Technology, Islamabad
Prof. Ashutosh Kumar Dubey, Trinity Institute of Technology and Research Bhopal, India
Mr. G. Appasami, Dr. Pauls Engineering College, India
Mr. M Yasin, National University of Science and Tech, karachi (NUST), Pakistan
Mr. Yaser Miaji, University Utara Malaysia, Malaysia
Mr. Shah Ahsanul Haque, International Islamic University Chittagong (IIUC), Bangladesh
Prof. (Dr) Syed Abdul Sattar, Royal Institute of Technology & Science, India
Dr. S. Sasikumar, Roever Engineering College
Assist. Prof. Monit Kapoor, Maharishi Markandeshwar University, India
Mr. Nwaocha Vivian O, National Open University of Nigeria
Dr. M. S. Vijaya, GR Govindarajulu School of Applied Computer Technology, India
Assist. Prof. Chakresh Kumar, Manav Rachna International University, India
Mr. Kunal Chadha , R&D Software Engineer, Gemalto, Singapore
Mr. Mueen Uddin, Universiti Teknologi Malaysia, UTM , Malaysia
Dr. Dhuha Basheer abdullah, Mosul university, Iraq
Mr. S. Audithan, Annamalai University, India
Prof. Vijay K Chaudhari, Technocrats Institute of Technology , India
Associate Prof. Mohd Ilyas Khan, Technocrats Institute of Technology , India
Dr. Vu Thanh Nguyen, University of Information Technology, HoChiMinh City, VietNam
Assist. Prof. Anand Sharma, MITS, Lakshmangarh, Sikar, Rajasthan, India
Prof. T V Narayana Rao, HITAM Engineering college, Hyderabad
Mr. Deepak Gour, Sir Padampat Singhania University, India
Assist. Prof. Amutharaj Joyson, Kalasalingam University, India
Mr. Ali Balador, Islamic Azad University, Iran
Mr. Mohit Jain, Maharaja Surajmal Institute of Technology, India
Mr. Dilip Kumar Sharma, GLA Institute of Technology & Management, India
Dr. Debojyoti Mitra, Sir padampat Singhania University, India
Dr. Ali Dehghantanha, Asia-Pacific University College of Technology and Innovation, Malaysia
Mr. Zhao Zhang, City University of Hong Kong, China
Prof. S.P. Setty, A.U. College of Engineering, India
Prof. Patel Rakeshkumar Kantilal, Sankalchand Patel College of Engineering, India
Mr. Biswajit Bhowmik, Bengal College of Engineering & Technology, India
Mr. Manoj Gupta, Apex Institute of Engineering & Technology, India
Assist. Prof. Ajay Sharma, Raj Kumar Goel Institute Of Technology, India
Assist. Prof. Ramveer Singh, Raj Kumar Goel Institute of Technology, India
Dr. Hanan Elazhary, Electronics Research Institute, Egypt
Dr. Hosam I. Faiq, USM, Malaysia
Prof. Dipti D. Patil, MAEER's MIT College of Engg. & Tech, Pune, India
Assist. Prof. Devendra Chack, BCT Kumaon engineering College Dwarahat Almora, India
Prof. Manpreet Singh, M. M. Engg. College, M. M. University, India
Assist. Prof. M. Sadiq ali Khan, University of Karachi, Pakistan
Mr. Prasad S. Halgaonkar, MIT - College of Engineering, Pune, India
Dr. Imran Ghani, Universiti Teknologi Malaysia, Malaysia
Prof. Varun Kumar Kakar, Kumaon Engineering College, Dwarahat, India

Assist. Prof. Nisheeth Joshi, Apaji Institute, Banasthali University, Rajasthan, India
Associate Prof. Kunwar S. Vaisla, VCT Kumaon Engineering College, India
Prof Anupam Choudhary, Bhilai School Of Engg.,Bhilai (C.G.),India
Mr. Divya Prakash Shrivastava, Al Jabal Al garbi University, Zawya, Libya
Associate Prof. Dr. V. Radha, Avinashilingam Deemed university for women, Coimbatore.
Dr. Kasarapu Ramani, JNT University, Anantapur, India
Dr. Anuraag Awasthi, Jayoti Vidyapeeth Womens University, India
Dr. C G Ravichandran, R V S College of Engineering and Technology, India
Dr. Mohamed A. Deriche, King Fahd University of Petroleum and Minerals, Saudi Arabia
Mr. Abbas Karimi, Universiti Putra Malaysia, Malaysia
Mr. Amit Kumar, Jaypee University of Engg. and Tech., India
Dr. Nikolai Stoianov, Defense Institute, Bulgaria
Assist. Prof. S. Ranichandra, KSR College of Arts and Science, Tiruchencode
Mr. T.K.P. Rajagopal, Diamond Horse International Pvt Ltd, India
Dr. Md. Ekramul Hamid, Rajshahi University, Bangladesh
Mr. Hemanta Kumar Kalita , TATA Consultancy Services (TCS), India
Dr. Messaouda Azzouzi, Ziane Achour University of Djelfa, Algeria
Prof. (Dr.) Juan Jose Martinez Castillo, "Gran Mariscal de Ayacucho" University and Acantelys research Group, Venezuela
Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India
Dr. Babak Bashari Rad, University Technology of Malaysia, Malaysia
Dr. Nighat Mir, Effat University, Saudi Arabia
Prof. (Dr.) G.M.Nasira, Sasurie College of Engineering, India
Mr. Varun Mittal, Gemalto Pte Ltd, Singapore
Assist. Prof. Mrs P. Banumathi, Kathir College Of Engineering, Coimbatore
Assist. Prof. Quan Yuan, University of Wisconsin-Stevens Point, US
Dr. Pranam Paul, Narula Institute of Technology, Agarpara, West Bengal, India
Assist. Prof. J. Ramkumar, V.L.B Janakiammal college of Arts & Science, India
Mr. P. Sivakumar, Anna university, Chennai, India
Mr. Md. Humayun Kabir Biswas, King Khalid University, Kingdom of Saudi Arabia
Mr. Mayank Singh, J.P. Institute of Engg & Technology, Meerut, India
HJ. Kamaruzaman Jusoff, Universiti Putra Malaysia
Mr. Nikhil Patrick Lobo, CADES, India
Dr. Amit Wason, Rayat-Bahra Institute of Engineering & Boi-Technology, India
Dr. Rajesh Shrivastava, Govt. Benazir Science & Commerce College, Bhopal, India
Assist. Prof. Vishal Bharti, DCE, Gurgaon
Mrs. Sunita Bansal, Birla Institute of Technology & Science, India
Dr. R. Sudhakar, Dr.Mahalingam college of Engineering and Technology, India
Dr. Amit Kumar Garg, Shri Mata Vaishno Devi University, Katra(J&K), India
Assist. Prof. Raj Gaurang Tiwari, AZAD Institute of Engineering and Technology, India
Mr. Hamed Taherdoost, Tehran, Iran
Mr. Amin Daneshmand Malayeri, YRC, IAU, Malayer Branch, Iran
Mr. Shantanu Pal, University of Calcutta, India
Dr. Terry H. Walcott, E-Promag Consultancy Group, United Kingdom
Dr. Ezekiel U OKIKE, University of Ibadan, Nigeria
Mr. P. Mahalingam, Caledonian College of Engineering, Oman
Dr. Mahmoud M. A. Abd Ellatif, Mansoura University, Egypt

Prof. Kunwar S. Vaisla, BCT Kumaon Engineering College, India
Prof. Mahesh H. Panchal, Kalol Institute of Technology & Research Centre, India
Mr. Muhammad Asad, Technical University of Munich, Germany
Mr. AliReza Shams Shafiqh, Azad Islamic university, Iran
Prof. S. V. Nagaraj, RMK Engineering College, India
Mr. Ashikali M Hasan, Senior Researcher, CelNet security, India
Dr. Adnan Shahid Khan, University Technology Malaysia, Malaysia
Mr. Prakash Gajanan Burade, Nagpur University/ITM college of engg, Nagpur, India
Dr. Jagdish B.Helonde, Nagpur University/ITM college of engg, Nagpur, India
Professor, Doctor BOUHORMA Mohammed, Univertsity Abdelmalek Essaadi, Morocco
Mr. K. Thirumalaivasan, Pondicherry Engg. College, India
Mr. Umbarkar Anantkumar Janardan, Walchand College of Engineering, India
Mr. Ashish Chaurasia, Gyan Ganga Institute of Technology & Sciences, India
Mr. Sunil Taneja, Kurukshetra University, India
Mr. Fauzi Adi Rafrastra, Dian Nuswantoro University, Indonesia
Dr. Yaduvir Singh, Thapar University, India
Dr. Ioannis V. Koskosas, University of Western Macedonia, Greece
Dr. Vasantha Kalyani David, Avinashilingam University for women, Coimbatore
Dr. Ahmed Mansour Manasrah, Universiti Sains Malaysia, Malaysia
Miss. Nazanin Sadat Kazazi, University Technology Malaysia, Malaysia
Mr. Saeed Rasouli Heikalabad, Islamic Azad University - Tabriz Branch, Iran
Assoc. Prof. Dhirendra Mishra, SVKM's NMIMS University, India
Prof. Shapor Zarei, UAE Inventors Association, UAE
Prof. B.Raja Sarath Kumar, Lenora College of Engineering, India
Dr. Bashir Alam, Jamia millia Islamia, Delhi, India
Prof. Anant J Umbarkar, Walchand College of Engg., India
Assist. Prof. B. Bharathi, Sathyabama University, India
Dr. Fokrul Alom Mazarbhuiya, King Khalid University, Saudi Arabia
Prof. T.S.Jeyali Laseeth, Anna University of Technology, Tirunelveli, India
Dr. M. Balraju, Jawahar Lal Nehru Technological University Hyderabad, India
Dr. Vijayalakshmi M. N., R.V.College of Engineering, Bangalore
Prof. Walid Moudani, Lebanese University, Lebanon
Dr. Saurabh Pal, VBS Purvanchal University, Jaunpur, India
Associate Prof. Suneet Chaudhary, Dehradun Institute of Technology, India
Associate Prof. Dr. Manuj Darbari, BBD University, India
Ms. Prema Selvaraj, K.S.R College of Arts and Science, India
Assist. Prof. Ms.S.Sasikala, KSR College of Arts & Science, India
Mr. Sukhvinder Singh Deora, NC Institute of Computer Sciences, India
Dr. Abhay Bansal, Amity School of Engineering & Technology, India
Ms. Sumita Mishra, Amity School of Engineering and Technology, India
Professor S. Viswanadha Raju, JNT University Hyderabad, India
Mr. Asghar Shahrzad Khashandarag, Islamic Azad University Tabriz Branch, India
Mr. Manoj Sharma, Panipat Institute of Engg. & Technology, India
Mr. Shakeel Ahmed, King Faisal University, Saudi Arabia
Dr. Mohamed Ali Mahjoub, Institute of Engineer of Monastir, Tunisia
Mr. Adri Jovin J.J., SriGuru Institute of Technology, India
Dr. Sukumar Senthilkumar, Universiti Sains Malaysia, Malaysia

Mr. Rakesh Bharati, Dehradun Institute of Technology Dehradun, India
Mr. Shervan Fekri Ershad, Shiraz International University, Iran
Mr. Md. Sufiqul Islam, Daffodil International University, Bangladesh
Mr. Mahmudul Hasan, Daffodil International University, Bangladesh
Prof. Mandakini Tayade, UIT, RGTU, Bhopal, India
Ms. Sarla More, UIT, RGTU, Bhopal, India
Mr. Tushar Hrishikesh Jaware, R.C. Patel Institute of Technology, Shirpur, India
Ms. C. Divya, Dr G R Damodaran College of Science, Coimbatore, India
Mr. Fahimuddin Shaik, Annamacharya Institute of Technology & Sciences, India
Dr. M. N. Giri Prasad, JNTUCE,Pulivendula, A.P., India
Assist. Prof. Chintan M Bhatt, Charotar University of Science And Technology, India
Prof. Sahista Machchhar, Marwadi Education Foundation's Group of institutions, India
Assist. Prof. Navnish Goel, S. D. College Of Enginnering & Technology, India
Mr. Khaja Kamaluddin, Sirt University, Sirt, Libya
Mr. Mohammad Zaidul Karim, Daffodil International, Bangladesh
Mr. M. Vijayakumar, KSR College of Engineering, Tiruchengode, India
Mr. S. A. Ahsan Rajon, Khulna University, Bangladesh
Dr. Muhammad Mohsin Nazir, LCW University Lahore, Pakistan
Mr. Mohammad Asadul Hoque, University of Alabama, USA
Mr. P.V.Sarathchand, Indur Institute of Engineering and Technology, India
Mr. Durgesh Samadhiya, Chung Hua University, Taiwan
Dr Venu Kuthadi, University of Johannesburg, Johannesburg, RSA
Dr. (Er) Jasvir Singh, Guru Nanak Dev University, Amritsar, Punjab, India
Mr. Jasmin Cosic, Min. of the Interior of Una-sana canton, B&H, Bosnia and Herzegovina
Dr S. Rajalakshmi, Botho College, South Africa
Dr. Mohamed Sarrab, De Montfort University, UK
Mr. Basappa B. Kodada, Canara Engineering College, India
Assist. Prof. K. Ramana, Annamacharya Institute of Technology and Sciences, India
Dr. Ashu Gupta, Apeejay Institute of Management, Jalandhar, India
Assist. Prof. Shaik Rasool, Shadan College of Engineering & Technology, India
Assist. Prof. K. Suresh, Annamacharya Institute of Tech & Sci. Rajampet, AP, India
Dr . G. Singaravel, K.S.R. College of Engineering, India
Dr B. G. Geetha, K.S.R. College of Engineering, India
Assist. Prof. Kavita Choudhary, ITM University, Gurgaon
Dr. Mehrdad Jalali, Azad University, Mashhad, Iran
Megha Goel, Shamli Institute of Engineering and Technology, Shamli, India
Mr. Chi-Hua Chen, Institute of Information Management, National Chiao-Tung University, Taiwan (R.O.C.)
Assoc. Prof. A. Rajendran, RVS College of Engineering and Technology, India
Assist. Prof. S. Jaganathan, RVS College of Engineering and Technology, India
Assoc. Prof. (Dr.) A S N Chakravarthy, JNTUK University College of Engineering Vizianagaram (State University)
Assist. Prof. Deepshikha Patel, Technocrat Institute of Technology, India
Assist. Prof. Maram Balajee, GMRIT, India
Assist. Prof. Monika Bhatnagar, TIT, India
Prof. Gaurang Panchal, Charotar University of Science & Technology, India
Prof. Anand K. Tripathi, Computer Society of India
Prof. Jyoti Chaudhary, High Performance Computing Research Lab, India
Assist. Prof. Supriya Raheja, ITM University, India

Dr. Pankaj Gupta, Microsoft Corporation, U.S.A.
Assist. Prof. Panchamukesh Chandaka, Hyderabad Institute of Tech. & Management, India
Prof. Mohan H.S, SJB Institute Of Technology, India
Mr. Hossein Malekinezhad, Islamic Azad University, Iran
Mr. Zatin Gupta, Universti Malaysia, Malaysia
Assist. Prof. Amit Chauhan, Phonics Group of Institutions, India
Assist. Prof. Ajal A. J., METS School Of Engineering, India
Mrs. Omowunmi Omobola Adeyemo, University of Ibadan, Nigeria
Dr. Bharat Bhushan Agarwal, I.F.T.M. University, India
Md. Nazrul Islam, University of Western Ontario, Canada
Tushar Kanti, L.N.C.T, Bhopal, India
Er. Aumreesh Kumar Saxena, SIRTs College Bhopal, India
Mr. Mohammad Monirul Islam, Daffodil International University, Bangladesh
Dr. Kashif Nisar, University Utara Malaysia, Malaysia
Dr. Wei Zheng, Rutgers Univ/ A10 Networks, USA
Associate Prof. Rituraj Jain, Vyas Institute of Engg & Tech, Jodhpur – Rajasthan
Assist. Prof. Apoorvi Sood, I.T.M. University, India
Dr. Kayhan Zrar Ghafoor, University Technology Malaysia, Malaysia
Mr. Swapnil Soner, Truba Institute College of Engineering & Technology, Indore, India
Ms. Yogita Gigras, I.T.M. University, India
Associate Prof. Neelima Sadineni, Pydha Engineering College, India Pydha Engineering College
Assist. Prof. K. Deepika Rani, HITAM, Hyderabad
Ms. Shikha Maheshwari, Jaipur Engineering College & Research Centre, India
Prof. Dr V S Giridhar Akula, Avanthi's Scientific Tech. & Research Academy, Hyderabad
Prof. Dr.S.Saravanan, Muthayammal Engineering College, India
Mr. Mehdi Golsorkhtabar Amiri, Islamic Azad University, Iran
Prof. Amit Sadanand Savyanavar, MITCOE, Pune, India
Assist. Prof. P.Oliver Jayaprakash, Anna University,Chennai
Assist. Prof. Ms. Sujata, ITM University, Gurgaon, India
Dr. Asoke Nath, St. Xavier's College, India
Mr. Masoud Rafighi, Islamic Azad University, Iran
Assist. Prof. RamBabu Pemula, NIMRA College of Engineering & Technology, India
Assist. Prof. Ms Rita Chhikara, ITM University, Gurgaon, India
Mr. Sandeep Maan, Government Post Graduate College, India
Prof. Dr. S. Muralidharan, Mepco Schlenk Engineering College, India
Associate Prof. T.V.Sai Krishna, QIS College of Engineering and Technology, India
Mr. R. Balu, Bharathiar University, Coimbatore, India
Assist. Prof. Shekhar. R, Dr.SM College of Engineering, India
Prof. P. Senthilkumar, Vivekananda Institue of Engineering and Techology for Woman, India
Mr. M. Kamarajan, PSNA College of Engineering & Technology, India
Dr. Angajala Srinivasa Rao, Jawaharlal Nehru Technical University, India
Assist. Prof. C. Venkatesh, A.I.T.S, Rajampet, India
Mr. Afshin Rezakhani Roozbahani, Ayatollah Boroujerdi University, Iran
Mr. Laxmi chand, SCTL, Noida, India
Dr. Dr. Abdul Hannan, Vivekanand College, Aurangabad
Prof. Mahesh Panchal, KITRC, Gujarat
Dr. A. Subramani, K.S.R. College of Engineering, Tiruchengode

Assist. Prof. Prakash M, Rajalakshmi Engineering College, Chennai, India
Assist. Prof. Akhilesh K Sharma, Sir Padampat Singhania University, India
Ms. Varsha Sahni, Guru Nanak Dev Engineering College, Ludhiana, India
Associate Prof. Trilochan Rout, NM Institute of Engineering and Technology, India
Mr. Srikanta Kumar Mohapatra, NMIET, Orissa, India
Mr. Waqas Haider Bangyal, Iqra University Islamabad, Pakistan
Dr. S. Vijayaragavan, Christ College of Engineering and Technology, Pondicherry, India
Prof. Elboukhari Mohamed, University Mohammed First, Oujda, Morocco
Dr. Muhammad Asif Khan, King Faisal University, Saudi Arabia
Dr. Nagy Ramadan Darwish Omran, Cairo University, Egypt.
Assistant Prof. Anand Nayyar, KCL Institute of Management and Technology, India
Mr. G. Premsankar, Ericsson, India
Assist. Prof. T. Hemalatha, VELS University, India
Prof. Tejaswini Apte, University of Pune, India
Dr. Edmund Ng Giap Weng, Universiti Malaysia Sarawak, Malaysia
Mr. Mahdi Nouri, Iran University of Science and Technology, Iran
Associate Prof. S. Asif Hussain, Annamacharya Institute of technology & Sciences, India
Mrs. Kavita Pabreja, Maharaja Surajmal Institute (an affiliate of GGSIP University), India
Mr. Vorugunti Chandra Sekhar, DA-IICT, India
Mr. Muhammad Najmi Ahmad Zabidi, Universiti Teknologi Malaysia, Malaysia
Dr. Aderemi A. Atayero, Covenant University, Nigeria
Assist. Prof. Osama Sohaib, Balochistan University of Information Technology, Pakistan
Assist. Prof. K. Suresh, Annamacharya Institute of Technology and Sciences, India
Mr. Hassen Mohammed Abdullaah Alsafi, International Islamic University Malaysia (IIUM) Malaysia
Mr. Robail Yasrab, Virtual University of Pakistan, Pakistan
Mr. R. Balu, Bharathiar University, Coimbatore, India
Prof. Anand Nayyar, KCL Institute of Management and Technology, Jalandhar
Assoc. Prof. Vivek S Deshpande, MIT College of Engineering, India
Prof. K. Saravanan, Anna university Coimbatore, India
Dr. Ravendra Singh, MJP Rohilkhand University, Bareilly, India
Mr. V. Mathivanan, IBRA College of Technology, Sultanate of OMAN
Assoc. Prof. S. Asif Hussain, AITS, India
Assist. Prof. C. Venkatesh, AITS, India
Mr. Sami Ulhaq, SZABIST Islamabad, Pakistan
Dr. B. Justus Rabi, Institute of Science & Technology, India
Mr. Anuj Kumar Yadav, Dehradun Institute of technology, India
Mr. Alejandro Mosquera, University of Alicante, Spain
Assist. Prof. Arjun Singh, Sir Padampat Singhania University (SPSU), Udaipur, India
Dr. Smriti Agrawal, JB Institute of Engineering and Technology, Hyderabad
Assist. Prof. Swathi Sambangi, Visakha Institute of Engineering and Technology, India
Ms. Prabhjot Kaur, Guru Gobind Singh Indraprastha University, India
Mrs. Samaher AL-Hothali, Yanbu University College, Saudi Arabia
Prof. Rajneeshkaur Bedi, MIT College of Engineering, Pune, India
Mr. Hassen Mohammed Abdullaah Alsafi, International Islamic University Malaysia (IIUM)
Dr. Wei Zhang, Amazon.com, Seattle, WA, USA
Mr. B. Santhosh Kumar, C S I College of Engineering, Tamil Nadu
Dr. K. Reji Kumar, , N S S College, Pandalam, India

Assoc. Prof. K. Seshadri Sastry, EIILM University, India
Mr. Kai Pan, UNC Charlotte, USA
Mr. Ruikar Sachin, SGGSIET, India
Prof. (Dr.) Vinodani Katiyar, Sri Ramswaroop Memorial University, India
Assoc. Prof., M. Giri, Sreenivasa Institute of Technology and Management Studies, India
Assoc. Prof. Labib Francis Gergis, Misr Academy for Engineering and Technology (MET), Egypt
Assist. Prof. Amanpreet Kaur, ITM University, India
Assist. Prof. Anand Singh Rajawat, Shri Vaishnav Institute of Technology & Science, Indore
Mrs. Hadeel Saleh Haj Aliwi, Universiti Sains Malaysia (USM), Malaysia
Dr. Abhay Bansal, Amity University, India
Dr. Mohammad A. Mezher, Fahad Bin Sultan University, KSA
Assist. Prof. Nidhi Arora, M.C.A. Institute, India
Prof. Dr. P. Suresh, Karpagam College of Engineering, Coimbatore, India
Dr. Kannan Balasubramanian, Mepco Schlenk Engineering College, India
Dr. S. Sankara Gomathi, Panimalar Engineering college, India
Prof. Anil Kumar Suthar, Gujarat Technological University, L.C. Institute of Technology, India
Assist. Prof. R. Hubert Rajan, NOORUL ISLAM UNIVERSITY, India
Assist. Prof. Dr. Jyoti Mahajan, College of Engineering & Technology
Assist. Prof. Homam Reda El-Taj, College of Network Engineering, Saudi Arabia & Malaysia
Mr. Bijan Paul, Shahjalal University of Science & Technology, Bangladesh
Assoc. Prof. Dr. Ch V Phani Krishna, KL University, India
Dr. Vishal Bhatnagar, Ambedkar Institute of Advanced Communication Technologies & Research, India
Dr. Lamri LAOUAMER, Al Qassim University, Dept. Info. Systems & European University of Brittany, Dept. Computer Science, UBO, Brest, France
Prof. Ashish Babanrao Sasankar, G.H.Raisoni Institute Of Information Technology, India
Prof. Pawan Kumar Goel, Shamli Institute of Engineering and Technology, India
Mr. Ram Kumar Singh, S.V Subharti University, India
Assistant Prof. Sunish Kumar O S, Amaljyothi College of Engineering, India
Dr Sanjay Bhargava, Banasthali University, India
Mr. Pankaj S. Kulkarni, AVEW's Shatabdi Institute of Technology, India
Mr. Roohollah Etemadi, Islamic Azad University, Iran
Mr. Oloruntoyin Sefiu Taiwo, Emmanuel Alayande College Of Education, Nigeria
Mr. Sumit Goyal, National Dairy Research Institute, India
Mr Jaswinder Singh Dilawari, Geeta Engineering College, India
Prof. Raghuraj Singh, Harcourt Butler Technological Institute, Kanpur
Dr. S.K. Mahendran, Anna University, Chennai, India
Dr. Amit Wason, Hindustan Institute of Technology & Management, Punjab
Dr. Ashu Gupta, Apeejay Institute of Management, India
Assist. Prof. D. Asir Antony Gnana Singh, M.I.E.T Engineering College, India
Mrs Mina Farmanbar, Eastern Mediterranean University, Famagusta, North Cyprus
Mr. Maram Balajee, GMR Institute of Technology, India
Mr. Moiz S. Ansari, Isra University, Hyderabad, Pakistan
Mr. Adebayo, Olawale Surajudeen, Federal University of Technology Minna, Nigeria
Mr. Jasvir Singh, University College Of Engg., India
Mr. Vivek Tiwari, MANIT, Bhopal, India
Assoc. Prof. R. Navaneethakrishnan, Bharathiyar College of Engineering and Technology, India
Mr. Somdip Dey, St. Xavier's College, Kolkata, India

Mr. Souleymane Balla-Arabé, Xi'an University of Electronic Science and Technology, China
Mr. Mahabub Alam, Rajshahi University of Engineering and Technology, Bangladesh
Mr. Sathyaprakash P., S.K.P Engineering College, India
Dr. N. Karthikeyan, SNS College of Engineering, Anna University, India
Dr. Binod Kumar, JSPM's, Jayawant Technical Campus, Pune, India
Assoc. Prof. Dinesh Goyal, Suresh Gyan Vihar University, India
Mr. Md. Abdul Ahad, K L University, India
Mr. Vikas Bajpai, The LNM IIT, India
Dr. Manish Kumar Anand, Salesforce (R & D Analytics), San Francisco, USA
Assist. Prof. Dheeraj Murari, Kumaon Engineering College, India
Assoc. Prof. Dr. A. Muthukumaravel, VELS University, Chennai
Mr. A. Siles Balasingh, St.Joseph University in Tanzania, Tanzania
Mr. Ravindra Daga Badgujar, R C Patel Institute of Technology, India
Dr. Preeti Khanna, SVKM's NMIMS, School of Business Management, India
Mr. Kumar Dayanand, Cambridge Institute of Technology, India
Dr. Syed Asif Ali, SMI University Karachi, Pakistan
Prof. Pallvi Pandit, Himachal Pradesh University, India
Mr. Ricardo Verschueren, University of Gloucestershire, UK
Assist. Prof. Mamta Juneja, University Institute of Engineering and Technology, Panjab University, India
Assoc. Prof. P. Surendra Varma, NRI Institute of Technology, JNTU Kakinada, India
Assist. Prof. Gaurav Shrivastava, RGPV / SVITS Indore, India
Dr. S. Sumathi, Anna University, India
Assist. Prof. Ankita M. Kapadia, Charotar University of Science and Technology, India
Mr. Deepak Kumar, Indian Institute of Technology (BHU), India
Dr. Dr. Rajan Gupta, GGSIP University, New Delhi, India
Assist. Prof M. Anand Kumar, Karpagam University, Coimbatore, India
Mr. Mr Arshad Mansoor, Pakistan Aeronautical Complex
Mr. Kapil Kumar Gupta, Ansal Institute of Technology and Management, India
Dr. Neeraj Tomer, SINE International Institute of Technology, Jaipur, India
Assist. Prof. Trunal J. Patel, C.G.Patel Institute of Technology, Uka Tarsadia University, Bardoli, Surat
Mr. Sivakumar, Codework solutions, India
Mr. Mohammad Sadegh Mirzaei, PGNR Company, Iran
Dr. Gerard G. Dumancas, Oklahoma Medical Research Foundation, USA
Mr. Varadala Sridhar, Varadhaman College Engineering College, Affiliated To JNTU, Hyderabad
Assist. Prof. Manoj Dhawan, SVITS, Indore
Assoc. Prof. Chitreshh Banerjee, Suresh Gyan Vihar University, Jaipur, India
Dr. S. Santhi, SCSVMV University, India
Mr. Davood Mohammadi Souran, Ministry of Energy of Iran, Iran
Mr. Shamim Ahmed, Bangladesh University of Business and Technology, Bangladesh
Mr. Sandeep Reddivari, Mississippi State University, USA
Assoc. Prof. Ousmane Thiare, Gaston Berger University, Senegal
Dr. Hazra Imran, Athabasca University, Canada
Dr. Setu Kumar Chaturvedi, Technocrats Institute of Technology, Bhopal, India
Mr. Mohd Dilshad Ansari, Jaypee University of Information Technology, India
Ms. Jaspreet Kaur, Distance Education LPU, India
Dr. D. Nagarajan, Salalah College of Technology, Sultanate of Oman
Dr. K.V.N.R.Sai Krishna, S.V.R.M. College, India

- Mr. Himanshu Pareek, Center for Development of Advanced Computing (CDAC), India
Mr. Khaldi Amine, Badji Mokhtar University, Algeria
Mr. Mohammad Sadegh Mirzaei, Scientific Applied University, Iran
Assist. Prof. Khyati Chaudhary, Ram-eesh Institute of Engg. & Technology, India
Mr. Sanjay Agal, Pacific College of Engineering Udaipur, India
Mr. Abdul Mateen Ansari, King Khalid University, Saudi Arabia
Dr. H.S. Behera, Veer Surendra Sai University of Technology (VSSUT), India
Dr. Shrikant Tiwari, Shri Shankaracharya Group of Institutions (SSGI), India
Prof. Ganesh B. Regulwar, Shri Shankarprasad Agnihotri College of Engg, India
Prof. Pinnamaneni Bhanu Prasad, Matrix vision GmbH, Germany
Dr. Shrikant Tiwari, Shri Shankaracharya Technical Campus (SSTC), India
Dr. Siddesh G.K., : Dayananada Sagar College of Engineering, Bangalore, India
Dr. Nadir Bouchama, CERIST Research Center, Algeria
Dr. R. Sathishkumar, Sri Venkateswara College of Engineering, India
Assistant Prof (Dr.) Mohamed Moussaoui, Abdelmalek Essaadi University, Morocco
Dr. S. Malathi, Panimalar Engineering College, Chennai, India
Dr. V. Subedha, Panimalar Institute of Technology, Chennai, India
Dr. Prashant Panse, Swami Vivekanand College of Engineering, Indore, India
Dr. Hamza Aldabbas, Al-Balqa'a Applied University, Jordan
Dr. G. Rasitha Banu, Vel's University, Chennai
Dr. V. D. Ambeth Kumar, Panimalar Engineering College, Chennai
Prof. Anuranjan Misra, Bhagwant Institute of Technology, Ghaziabad, India
Ms. U. Sinthuja, PSG college of arts &science, India
Dr. Ehsan Saradar Torshizi, Urmia University, Iran
Dr. Shamneesh Sharma, APG Shimla University, Shimla (H.P.), India
Assistant Prof. A. S. Syed Navaz, Muthayammal College of Arts & Science, India
Assistant Prof. Ranjit Panigrahi, Sikkim Manipal Institute of Technology, Majitar, Sikkim
Dr. Khaled Eskaf, Arab Academy for Science ,Technology & Maritime Transportation, Egypt
Dr. Nishant Gupta, University of Jammu, India
Assistant Prof. Nagarajan Sankaran, Annamalai University, Chidambaram, Tamilnadu, India
Assistant Prof. Tribikram Pradhan, Manipal Institute of Technology, India
Dr. Nasser Lotfi, Eastern Mediterranean University, Northern Cyprus
Dr. R. Manavalan, K S Rangasamy college of Arts and Science, Tamilnadu, India
Assistant Prof. P. Krishna Sankar, K S Rangasamy college of Arts and Science, Tamilnadu, India
Dr. Rahul Malik, Cisco Systems, USA
Dr. S. C. Lingareddy, ALPHA College of Engineering, India
Assistant Prof. Mohammed Shuaib, Interl University, Lucknow, India
Dr. Sachin Yele, Sanghvi Institute of Management & Science, India
Dr. T. Thambidurai, Sun Univercell, Singapore
Prof. Anandkumar Telang, BKIT, India
Assistant Prof. R. Poorvadevi, SCVMV University, India
Dr Uttam Mande, Gitam University, India
Dr. Poornima Girish Naik, Shahu Institute of Business Education and Research (SIBER), India
Prof. Md. Abu Kausar, Jaipur National University, Jaipur, India
Dr. Mohammed Zuber, AISECT University, India
Prof. Kalum Priyanath Udagepola, King Abdulaziz University, Saudi Arabia
Dr. K. R. Ananth, Velalar College of Engineering and Technology, India

Assistant Prof. Sanjay Sharma, Roorkee Engineering & Management Institute Shamli (U.P), India
Assistant Prof. Panem Charan Arur, Priyadarshini Institute of Technology, India
Dr. Ashwak Mahmood muhsen alabaichi, Karbala University / College of Science, Iraq
Dr. Urmila Shrawankar, G H Raisoni College of Engineering, Nagpur (MS), India
Dr. Krishan Kumar Paliwal, Panipat Institute of Engineering & Technology, India
Dr. Mukesh Negi, Tech Mahindra, India
Dr. Anuj Kumar Singh, Amity University Gurgaon, India
Dr. Babar Shah, Gyeongsang National University, South Korea
Assistant Prof. Jayprakash Upadhyay, SRI-TECH Jabalpur, India
Assistant Prof. Varadala Sridhar, Vidya Jyothi Institute of Technology, India
Assistant Prof. Parameshachari B D, KSIT, Bangalore, India
Assistant Prof. Ankit Garg, Amity University, Haryana, India
Assistant Prof. Rajashe Karappa, SDMCET, Karnataka, India
Assistant Prof. Varun Jasuja, GNIT, India
Assistant Prof. Sonal Honale, Abha Gaikwad Patil College of Engineering Nagpur, India
Dr. Pooja Choudhary, CT Group of Institutions, NIT Jalandhar, India
Dr. Faouzi Hidoussi, UHL Batna, Algeria
Dr. Naseer Ali Husieen, Wasit University, Iraq
Assistant Prof. Vinod Kumar Shukla, Amity University, Dubai
Dr. Ahmed Farouk Metwaly, K L University
Mr. Mohammed Noaman Murad, Cihan University, Iraq
Dr. Suxing Liu, Arkansas State University, USA
Dr. M. Gomathi, Velalar College of Engineering and Technology, India
Assistant Prof. Sumardiono, College PGRI Blitar, Indonesia
Dr. Latika Kharb, Jagan Institute of Management Studies (JIMS), Delhi, India
Associate Prof. S. Raja, Pauls College of Engineering and Technology, Tamilnadu, India
Assistant Prof. Seyed Reza Pakize, Shahid Sani High School, Iran
Dr. Thiyyagu Nagaraj, University-INOU, India
Assistant Prof. Noreen Sarai, Harare Institute of Technology, Zimbabwe
Assistant Prof. Gajanand Sharma, Suresh Gyan Vihar University Jaipur, Rajasthan, India
Assistant Prof. Mapari Vikas Prakash, Siddhant COE, Sudumbare, Pune, India
Dr. Devesh Katiyar, Shri Ramswaroop Memorial University, India
Dr. Shenshen Liang, University of California, Santa Cruz, US
Assistant Prof. Mohammad Abu Omar, Limkokwing University of Creative Technology- Malaysia
Mr. Snehasis Banerjee, Tata Consultancy Services, India
Assistant Prof. Kibona Lusekelo, Ruaha Catholic University (RUCU), Tanzania
Assistant Prof. Adib Kabir Chowdhury, University College Technology Sarawak, Malaysia
Dr. Ying Yang, Computer Science Department, Yale University, USA
Dr. Vinay Shukla, Institute Of Technology & Management, India
Dr. Liviu Octavian Mafteiu-Scai, West University of Timisoara, Romania
Assistant Prof. Rana Khudhair Abbas Ahmed, Al-Rafidain University College, Iraq
Assistant Prof. Nitin A. Naik, S.R.T.M. University, India
Dr. Timothy Powers, University of Hertfordshire, UK
Dr. S. Prasath, Bharathiar University, Erode, India
Dr. Ritu Shrivastava, SIRTS Bhopal, India
Prof. Rohit Shrivastava, Mittal Institute of Technology, Bhopal, India
Dr. Gianina Mihai, Dunarea de Jos" University of Galati, Romania

Assistant Prof. Ms. T. Kalai Selvi, Erode Sengunthar Engineering College, India
Assistant Prof. Ms. C. Kavitha, Erode Sengunthar Engineering College, India
Assistant Prof. K. Sinivasamoorthi, Erode Sengunthar Engineering College, India
Assistant Prof. Mallikarjun C Sarsamba Bheemnna Khandre Institute Technology, Bhalki, India
Assistant Prof. Vishwanath Chikaraddi, Veermata Jijabai technological Institute (Central Technological Institute), India
Assistant Prof. Dr. Ikvinderpal Singh, Trai Shatabdi GGS Khalsa College, India
Assistant Prof. Mohammed Noaman Murad, Cihan University, Iraq
Professor Yousef Farhaoui, Moulay Ismail University, Errachidia, Morocco
Dr. Parul Verma, Amity University, India
Professor Yousef Farhaoui, Moulay Ismail University, Errachidia, Morocco
Assistant Prof. Madhavi Dhingra, Amity University, Madhya Pradesh, India
Assistant Prof.. G. Selvavinayagam, SNS College of Technology, Coimbatore, India
Assistant Prof. Madhavi Dhingra, Amity University, MP, India
Professor Kartheesan Log, Anna University, Chennai
Professor Vasudeva Acharya, Shri Madhwa vadira Institute of Technology, India
Dr. Asif Iqbal Hajamydeen, Management & Science University, Malaysia
Assistant Prof., Mahendra Singh Meena, Amity University Haryana
Assistant Professor Manjeet Kaur, Amity University Haryana
Dr. Mohamed Abd El-Basset Matwalli, Zagazig University, Egypt
Dr. Ramani Kannan, Universiti Teknologi PETRONAS, Malaysia
Assistant Prof. S. Jagadeesan Subramaniam, Anna University, India
Assistant Prof. Dharmendra Choudhary, Tripura University, India
Assistant Prof. Deepika Vodnala, SR Engineering College, India
Dr. Kai Cong, Intel Corporation & Computer Science Department, Portland State University, USA
Dr. Kailas R Patil, Vishwakarma Institute of Information Technology (VIIT), India
Dr. Omar A. Alzubi, Faculty of IT / Al-Balqa Applied University, Jordan
Assistant Prof. Kareemullah Shaik, Nimra Institute of Science and Technology, India
Assistant Prof. Chirag Modi, NIT Goa
Dr. R. Ramkumar, Nandha Arts And Science College, India
Dr. Priyadarshini Vydhalingam, Harathiar University, India
Dr. P. S. Jagadeesh Kumar, DBIT, Bangalore, Karnataka
Dr. Vikas Thada, AMITY University, Pachgaon
Dr. T. A. Ashok Kumar, Institute of Management, Christ University, Bangalore
Dr. Shaheera Rashwan, Informatics Research Institute
Dr. S. Preetha Gunasekar, Bharathiyar University, India
Asst Professor Sameer Dev Sharma, Uttarakhand University, Dehradun
Dr. Zhihan Lv, Chinese Academy of Science, China
Dr. Ikvinderpal Singh, Trai Shatabdi GGS Khalsa College, Amritsar
Dr. Umar Ruhi, University of Ottawa, Canada
Dr. Jasmin Cosic, University of BiHac, Bosnia and Herzegovina
Dr. Homam Reda El-Taj, University of Tabuk, Kingdom of Saudi Arabia
Dr. Mostafa Ghobaei Arani, Islamic Azad University, Iran
Dr. Ayyasamy Ayyanar, Annamalai University, India
Dr. Selvakumar Manickam, Universiti Sains Malaysia, Malaysia
Dr. Murali Krishna Namana, GITAM University, India
Dr. Smriti Agrawal, Chaitanya Bharathi Institute of Technology, Hyderabad, India
Professor Vimalathithan Rathinasabapathy, Karpagam College Of Engineering, India

Dr. Sushil Chandra Dimri, Graphic Era University, India
Dr. Dinh-Sinh Mai, Le Quy Don Technical University, Vietnam
Dr. S. Rama Sree, Aditya Engg. College, India
Dr. Ehab T. Alnfrawy, Sadat Academy, Egypt
Dr. Patrick D. Cerna, Haramaya University, Ethiopia
Dr. Vishal Jain, Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), India
Associate Prof. Dr. Jiliang Zhang, North Eastern University, China
Dr. Sharefa Murad, Middle East University, Jordan
Dr. Ajeet Singh Poonia, Govt. College of Engineering & technology, Rajasthan, India
Dr. Vahid Esmaeelzadeh, University of Science and Technology, Iran
Dr. Jacek M. Czerniak, Casimir the Great University in Bydgoszcz, Institute of Technology, Poland
Associate Prof. Anisur Rehman Nasir, Jamia Millia Islamia University
Assistant Prof. Imran Ahmad, COMSATS Institute of Information Technology, Pakistan
Professor Ghulam Qasim, Preston University, Islamabad, Pakistan
Dr. Parameshachari B D, GSSS Institute of Engineering and Technology for Women
Dr. Wencan Luo, University of Pittsburgh, US
Dr. Musa PEKER, Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Gunasekaran Shanmugam, Anna University, India
Dr. Binh P. Nguyen, National University of Singapore, Singapore
Dr. Rajkumar Jain, Indian Institute of Technology Indore, India
Dr. Imtiaz Ali Halepoto, QUEST Nawabshah, Pakistan
Dr. Shaligram Prajapat, Devi Ahilya University Indore India
Dr. Sunita Singhal, Birla Institute of Technology and Science, Pilani, India
Dr. Ijaz Ali Shoukat, King Saud University, Saudi Arabia
Dr. Anuj Gupta, IKG Punjab Technical University, India
Dr. Sonali Saini, IES-IPS Academy, India
Dr. Krishan Kumar, MotiLal Nehru National Institute of Technology, Allahabad, India
Dr. Z. Faizal Khan, College of Engineering, Shaqra University, Kingdom of Saudi Arabia
Prof. M. Padmavathamma, S.V. University Tirupati, India
Prof. A. Velayudham, Cape Institute of Technology, India
Prof. Seifeide Kadry, American University of the Middle East
Dr. J. Durga Prasad Rao, Pt. Ravishankar Shukla University, Raipur
Assistant Prof. Najam Hasan, Dhofar University
Dr. G. Suseendran, Vels University, Pallavaram, Chennai
Prof. Ankit Faldu, Gujarat Technological University- Atmiya Institute of Technology and Science
Dr. Ali Habiboghi, Islamic Azad University
Dr. Deepak Dembla, JECRC University, Jaipur, India
Dr. Pankaj Rajan, Walmart Labs, USA
Assistant Prof. Radoslava Kraleva, South-West University "Neofit Rilski", Bulgaria
Assistant Prof. Medhavi Shriwas, Shri Vaishnav Institute of Technology, India
Associate Prof. Sedat Akleylek, Ondokuz Mayis University, Turkey
Dr. U.V. Arivazhagu, Kingston Engineering College Affiliated To Anna University, India
Dr. Touseef Ali, University of Engineering and Technology, Taxila, Pakistan
Assistant Prof. Naren Jeeva, SASTRA University, India
Dr. Riccardo Colella, University of Salento, Italy
Dr. Enache Maria Cristina, University of Galati, Romania
Dr. Senthil P, Kurinji College of Arts & Science, India

- Dr. Hasan Ashrafi-rizi, Isfahan University of Medical Sciences, Isfahan, Iran
Dr. Mazhar Malik, Institute of Southern Punjab, Pakistan
Dr. Yajie Miao, Carnegie Mellon University, USA
Dr. Kamran Shaukat, University of the Punjab, Pakistan
Dr. Sasikaladevi N., SASTRA University, India
Dr. Ali Asghar Rahmani Hosseiniabadi, Islamic Azad University Ayatollah Amoli Branch, Amol, Iran
Dr. Velin Kralev, South-West University "Neofit Rilski", Blagoevgrad, Bulgaria
Dr. Marius Iulian Mihailescu, LUMINA - The University of South-East Europe
Dr. Sriramula Nagaprasad, S.R.R.Govt.Arts & Science College, Karimnagar, India
Prof (Dr.) Namrata Dhanda, Dr. APJ Abdul Kalam Technical University, Lucknow, India
Dr. Javed Ahmed Maher, Shah Abdul Latif University, Khairpur Mir's, Pakistan
Dr. B. Narendra Kumar Rao, Sree Vidyanikethan Engineering College, India
Dr. Shahzad Anwar, University of Engineering & Technology Peshawar, Pakistan
Dr. Basit Shahzad, King Saud University, Riyadh - Saudi Arabia
Dr. Nilamadhab Mishra, Chang Gung University
Dr. Sachin Kumar, Indian Institute of Technology Roorkee
Dr. Santosh Nanda, Biju-Patnaik University of Technology
Dr. Sherzod Turaev, International Islamic University Malaysia
Dr. Yilun Shang, Tongji University, Department of Mathematics, Shanghai, China
Dr. Nuzhat Shaikh, Modern Education society's College of Engineering, Pune, India
Dr. Parul Verma, Amity University, Lucknow campus, India
Dr. Rachid Alaoui, Agadir Ibn Zohr University, Agadir, Morocco
Dr. Dharmendra Patel, Charotar University of Science and Technology, India
Dr. Dong Zhang, University of Central Florida, USA
Dr. Kennedy Chinedu Okafor, Federal University of Technology Owerri, Nigeria
Prof. C Ram Kumar, Dr NGP Institute of Technology, India
Dr. Sandeep Gupta, GGS IP University, New Delhi, India
Dr. Shahanawaj Ahamad, University of Ha'il, Ha'il City, Ministry of Higher Education, Kingdom of Saudi Arabia
Dr. Najeed Ahmed Khan, NED University of Engineering & Technology, India
Dr. Sajid Ullah Khan, Universiti Malaysia Sarawak, Malaysia
Dr. Muhammad Asif, National Textile University Faisalabad, Pakistan
Dr. Yu BI, University of Central Florida, Orlando, FL, USA
Dr. Brijendra Kumar Joshi, Research Center, Military College of Telecommunication Engineering, India
Prof. Dr. Nak Eun Cho, Pukyong National University, Korea
Prof. Wasim Ul-Haq, Faculty of Science, Majmaah University, Saudi Arabia
Dr. Mohsan Raza, G.C University Faisalabad, Pakistan
Dr. Syed Zakar Hussain Bukhari, National Science and Technalogy Azad Jamu Kashmir, Pakistan
Dr. Ruksar Fatima, KBN College of Engineering, Gulbarga, Karnataka, India
Associate Professor S. Karpagavalli, Department of Computer Science, PSGR Krishnammal College for Women
Coimbatore, Tamilnadu, India
Dr. Bushra Mohamed Elamin Elnaim, Prince Sattam bin Abdulaziz University, Saudi Arabia
Dr. Shamik Tiwari, Department of CSE, CET, Mody University, Lakshmangarh
Dr. Rohit Raja, Faculty of Engineering and Technology, Shri Shankaracharya Group of Institutions, India
Prof. Dr. Aqeel-ur-Rehman, Department of Computing, HIET, FEST, Hamdard University, Pakistan
Dr. Nageswara Rao Moparthi, Velagapudi Ramakrishna Siddhartha Engineering College, India
Dr. Mohd Muqeem, Department of Computer Application, Integral University, Lucknow, India
Dr. Zeeshan Bhatti, Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan

- Dr. Emrah Irmak, Biomedical Engineering Department, Karabuk University, Turkey
Dr. Fouad Abdulameer salman, School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu
Dr. N. Prasath, Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Arasur, Coimbatore
Dr. Hasan Ashrafi-rizi, Health Information Technology Research Center, Isfahan University of Medical Sciences, Hezar Jerib Avenue, Isfahan, Iran
Dr. N. Sasikaladevi, School of Computing, SASTRA University, Thirumalisamudram, Tamilnadu, India.
Dr. Anchit Bijalwan, Arba Minch University, Ethiopia
Dr. K. Sathishkumar, BlueCrest University College, Accra North, Ghana, West Africa
Dr. Dr. Parameshachari B D, GSSS Institute of Engineering and Technology for Women, Affiliated to Visvesvaraya Technological University, Belagavi
Dr. C. Shoba Bindu, Dept. of CSE, JNTUA College of Engineering, India
Dr. M. Inbavalli, ER. Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India
Dr. Vidya Sagar Ponnam, Dept. of IT, Velagapudi Ramakrishna Siddhartha Engineering College, India
Dr. Kelvin LO M. F., The Hong Kong Polytechnic University, Hong Kong
Prof. Karimella Vikram, G.H. Raisoni College of Engineering & Management, Pune, India
Dr. Shajilin Loret J.B., VV College of Engineering, India
Dr. P. Sujatha, Department of Computer Science at Vels University, Chennai
Dr. Vaibhav Sundriyal, Old Dominion University Research Foundation, USA
Dr. Md Masud Rana, Khulna University of Engineering and Technology, Bangladesh
Dr. Gurcharan Singh, Khasa College Amritsar, Guru Nanak Dev University, Amritsar, India
Dr. Richard Otieno Omollo, Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, Kenya
Prof. (Dr) Amit Verma, Computer Science & Engineering, Chandigarh Engineering College, Landran, Mohali, India
Dr. Vidya Sagar Ponnam, Velagapudi Ramakrishna Siddhartha Engineering College, India
Dr. Bohui Wang, School of Aerospace Science and Technology, Xidian University, P.R. China
Dr. M. Anjan Kumar, Department of Computer Science, Satavahana University, Karimnagar
Dr. Hanumanthappa J., DoS in CS, Uni of Mysuru, Karnataka, India
Dr. Pouya Derakhshan-Barjoei, Dept. of Telecommunication and Engineering, Islamic Azad University, Iran
Professor Edelberto Silva, Universidade Federal de Juiz de Fora, Brazil
Dr. Sonali Vyas, Amity University Rajasthan, India
Dr. Santosh Bharti, National Institute of Technology Rourkela, India
Dr. Deepak Gupta, Maharaja Agrasen Institute of Technology, India
Dr. Emrah Irmak, Karabuk University, Turkey
Dr. Yojna Arora, Amity University, India
Dr. Marta Cimitile, Uniteima Sapienza, Italy
Assistant Prof. Shanthakumari Raju, Kongu Engineering College, India
Dr. Ravi Verma, RGPV Bhopal, India
Dr. Tanweer Alam, Islamic University of Madinah, Dept. of Computer Science, College of Computer and Information System, Al Madinah, Saudi Arabia
Dr. Kumar Keshamoni, Dept. of ECE, Vaagdevi Engineering College, Warangal, Telangana, India
Dr. G. Rajkumar, N.M.S.S.Vellaichamy Nadar College, Madurai, Tamilnadu, India
Dr. P. Mayil Vel Kumar, Karpagam Institute of Technology, Coimbatore, India
Dr. M. Yaswanth Bhanu Murthy, Vasireddy Venkatadri Institute of Technology, Guntur, A.P., India
Asst. Prof. Dr. Mehmet Barış TABAKCIOĞLU, Bursa Technical University, Turkey
Dr. Mohd. Muntjir, College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia
Dr. Sanjay Agal, Aravali Institute of Technical Studies, Udaipur, India

Dr. Shanshan Tuo, xAd Inc., US
Dr. Subhadra Shaw, AKS University, Satna, India
Dr. Piyush Anand, Noida International University, Greater Noida, India
Dr. Brijendra Kumar Joshi, Research Center Military College of Telecommunication Engineering, India
Dr. V. Sreerama Murthy, GMRIT, Rajam, AP, India
Dr. S. Nagarajan, Annamalai University, India
Prof. Pramod Bhausaheb Deshmukh, D. Y. Patil College of Engineering, Akurdi, Pune, India
Dr. Jaspreet Kour, GCET, India
Dr. Parul Agarwal, Jamia Hamdard
Dr. Muhammad Faheem, Abdullaah Gul University
Dr. Vaibhav Sundriyal, Old Dominion University
Dr. Sujatha Dandu, JNTUH
Dr. Wenzhao Zhang, NCSU, US
Dr. Senthil Kumar P., Anna University
Dr. Harshal Karande, Arvind Gavali College of Engineering, Satara
Dr. Kannan Dhandapani, Nehru Arts and Science College, Affiliated to Bharatiar University
Prof. Dr. Muthukumar Subramanian, Indian Institute of Information Technology, Tamilnadu, India
Dr. K .Vengatesan Krishnasamy, Dr. BATU University
Dr. Jayapandian N., Knowledge Institute of Technology
Dr. Sangeetha S.K.B, Rajalakshmi Engineering College
Dr. Geetha Devi Appari, PVP Siddhartha Institute of Technology
Dr. Pradeep Gurunathan, A.V.C. College of Engineering
Dr. Muftah Fraifer, Interaction design Center-University of Limerick
Dr. Gamal Eladl, Mansoura University/ IS Dept.
Dr. Bereket Assa, Woliya Soddo University
Dr. Venkata Suryanarayana Tinnaluri, Malla Reddy Group of Institutions
Dr. Jagadeesh Gopal, VIT University, Vellore
Dr. Vidya Sagar Ponnam, JNTUK, Kakinada/Velagapudi Ramakrishna Siddhartha Engineering College
Dr. Meenashi Sharma, Chandigarh University
Dr. Hiyam Hatem, University of Baghdad, College of Science
Dr. Smitha Elsa Peter, PRIST University
Dr. Gurcharan Singh, Guru Nanak Dev University
Dr. Ahmed EL-YAHYAOUI, Mohammed V University in Rabat
Dr. Shruti Bahrgava, JNTUH
Dr. Seda Kul, Kocaeli University
Dr. Bappaditya Jana, Chaibasa Engineering College
Dr. Farhad Goodarzi, UPM university
Dr. Sujatha P., Vels University, Chennai
Dr. Satya Bhushan Verma, National Institute of Technology Durgapur
Dr. Man Fung LO, The Hong Kong Polytechnic University
Dr. Muhammad Adnan, Abdul Wali Khan University
Dr. Seyed Sahand Mohammadi Ziabari, Vrije University
Dr. Brindha Srinivasan, Palanisamy College of Arts, Erode
Dr. Mohammad Aldabbagh, University of Mosul
Prof. Abdallah Rhatto, Moulay Ismail University, Higher School of Technology
Dr. Kumar Keshamoni, Vaagdevi Engineering College, Warangal, Telangana, India
Dr. Khalid Nazim Abdus Sattar, College of Science, Az-Zulfi campus, Majmaah university, Kingdom of Saudi Arabia

CALL FOR PAPERS

International Journal of Computer Science and Information Security

IJCSIS 2018-2019

ISSN: 1947-5500

<http://sites.google.com/site/ijcsis/>

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc.,) Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on

its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embedded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at <http://sites.google.com/site/ijcsis/authors-notes>.

**© IJCSIS PUBLICATION 2018
ISSN 1947 5500
<http://sites.google.com/site/ijcsis/>**